A tutorial for conducting causal mediation analysis with the `twangMediation` package

Donna L. Coffman, Megan S. Schuler, Daniel F. McCaffrey,
Katherine E. Castellano, Haoyu Zhou, Brian Vegetabile, and Beth Ann Griffin

June 21, 2022

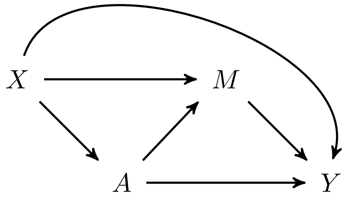
1 Introduction

The `twangMediation` R package is an extension of the Toolkit for Weighting and Analysis of Nonequivalent Groups (`twang`) R package that contains a set of functions to support causal modeling of observational data through the estimation and evaluation of propensity scores and propensity score-based weights. Currently, `twang` can be used to estimate treatment effects with two or more treatment groups and time-varying treatments. The `twangMediation` package builds on the `twang` package to estimate mediation effects for binary, ordinal, multinomial (categorical), or continuous mediator(s) of a binary exposure variable. This tutorial provides an introduction to causal mediation analysis using `twangMediation` and demonstrates its use through an illustrative example. We first provide a brief overview of causal mediation, including definitions of the natural direct and indirect estimands of interest, as well as the required identification assumptions. If you are already familiar with causal mediation, you can skip to Section 2.1 for an introduction to our illustrative example and to Section 5 for step-by-step instructions for the `twangMediation` functions for estimating causal mediation effects.

2 An Overview of Causal Mediation

An important scientific goal in many fields of research is determining to what extent the total effect of an exposure on an outcome is mediated by an intermediate variable on the causal pathway between the exposure and outcome. A simple mediation model is illustrated in Figure 1 where $Y \equiv$ outcome, $A \equiv$ exposure, $X \equiv$ pre-exposure covariates, and $M \equiv$ mediator. Note that we use "exposure" broadly to refer to a non-randomized or randomized condition, treatment, or intervention.

Figure 1: Graphical depiction of a simple mediation model.



The **total effect** of A on Y includes two possible causal paths from A to Y: the path $A \to M \to Y$ is the **indirect effect** of A on Y through M and the path $A \to Y$ is the **direct effect** of A on Y that does not go through M. Direct and indirect effects are of scientific interest because they provide a framework to quantify and characterize the mechanism by which an exposure affects a given outcome.

Traditionally, direct and indirect effects have been evaluated using linear model specifications for the observed data, assuming no interactions or nonlinearities involving A and M. The definitions of the direct and indirect effects themselves rely on this linear specification. In response, a fast-growing literature in causal inference focuses on the definition, identification, and estimation of direct and indirect effects in fully non-parametric models (i.e., does not rely on a linear model specification) primarily based on ideas developed by Robins and Greenland (1992) and Pearl (2001). These developments use potential outcomes/counterfactuals to give non-parametric definitions of the effects involved in mediation analysis, known as controlled direct effects, natural direct and indirect effects, and interventional effects. For an introduction to all of these effects, see Nguyen et. al. (2020). Here, we focus on the natural (in)direct effects.

Mediation is inherently about **causal** effects, which are defined as the difference between two potential outcomes for an individual. We begin by introducing the potential outcomes needed to define the natural direct and indirect effects.

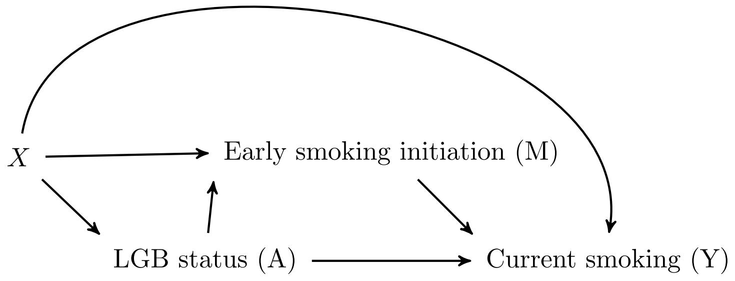
Consider the case in which A is a binary indicator of the exposure, indicating the exposed condition ($A = 1$) or the comparison condition ($A = 0$). There are two potential outcomes for each study participant corresponding to each exposure level a: the outcome had they received the exposure, denoted Y_1, and the outcome had they received the comparison condition, denoted Y_0. These two potential outcomes, Y_1 and Y_0, exist for all individuals in the population regardless of whether the individual received the exposure or comparison condition. However, we can observe only one of these outcomes for each participant depending on which exposure condition the individual actually receives.

The mediator is an "intermediate" outcome of the exposure and itself has potential values. For each exposure level a there is a corresponding potential mediator value, denoted M_a. Also, there is a corresponding potential outcome that reflects the outcome value that would arise under the specific exposure level a and the specific potential mediator value M_a – this potential outcome is denoted $Y_{(a,M_a)}$. Causal definitions of direct and indirect effects require extending the potential outcomes framework such that there is a potential outcome for each treatment and mediator pair. For the case of a binary exposure A, there are four potential outcomes for an individual, formed by crossing both potential exposure values with both potential mediator values: $Y_{(1,M_1)}$, $Y_{(0,M_0)}$, $Y_{(1,M_0)}$, and $Y_{(0,M_1)}$. Only $Y_{(1,M_1)}$ or $Y_{(0,M_0)}$, which correspond to the individual receiving $A = 1$ or $A = 0$ respectively, can be observed in practice. The other two potential outcomes are hypothetical quantities (i.e., the mediator value is manipulated to take on the value it would have under the other exposure condition); these are necessary to define the causal estimands of interest, as we detail later. Furthermore, for a given individual i, we can observe only one outcome, namely that which corresponds to the exposure level a that the individual received: $Y_{i,(A_i=a,M_{i,a}=m)}$. Before defining the natural direct and indirect effect estimands, we introduce our motivating example so that we may use it to more concretely define these effects.

2.1 Motivating Example

Our motivating example applies mediation analysis to health disparities research . Our specific focus is examining potential mediating pathways that explain substance use disparities among sexual minority (e.g., gay, lesbian, or bisexual) women, using data from the National Survey of Drug Use and Health (NSDUH). Specifically, lesbian, gay, and bisexual (LGB) women report higher rates of smoking and alcohol use than heterosexual women. We conceptualize sexual minority status as the exposure of interest, in that it gives rise to experiences of "minority stress," namely excess social stressors experienced by individuals in a marginalized social group (e.g., LGB individuals). Manifestations of minority stress may include experiences of stigma, discrimination, bullying, and family rejection, among others. Substance use among LGB individuals has been theorized to reflect, in part, a coping strategy to minority stress experiences. In our example, the particular outcome of interest is current smoking among LGB women, which we know to be disproportionately higher than among heterosexual women (Schuler & Collins, 2019). We apply mediation analysis to elucidate potential causal pathways that may give rise to these elevated rates of smoking. Specifically, our hypothesized mediator is early smoking initiation (i.e., prior to age 15); that is, we hypothesize that LGB girls are more likely to begin smoking at an early age than heterosexual women, potentially in response to minority stressors. Resultantly, early smoking initiation, which is a strong risk factor for developing nicotine dependence, contributes to higher rates of smoking among LGB women. In summary, the exposure is defined as sexual minority status (1=LGB women, 0=heterosexual women), the mediator is early smoking initiation (1=early initiation, 0=no early initiation), and the outcome is current smoking in adulthood (1=yes, 0=no). Baseline covariates include age, race/ethnicity, education level, household income, employment status, marital status, and urban vs. rural residence. Figure 2 illustrates our motivating example:

Figure 2: Graphical depiction of the effect of LGB status on adult smoking status as mediated by early smoking initiation.



2.2 Estimands: Natural direct and indirect effects

Causal effects are defined as contrasts between different potential outcomes. Specifically, our causal estimands of interest are the natural direct and natural indirect effects, defined below. First, we define the potential outcomes in the context of our motivating example. We consider two possible exposure values: LGB status, $A = 1$, and heterosexual status, $A = 0$ (note that these groups reflect the measurement of sexual identity in the NSDUH; individuals may identify as a broader range of sexual identities). Correspondingly, there are two potential mediator values: early smoking initiation status corresponding to LGB status, M_1, and early smoking initiation status corresponding to heterosexual status, M_0.

When we cross the possible exposure values and potential mediator values, there are four potential outcome values:

- $Y_{(1,M_1)}$, the potential outcome for adult smoking status when an individual is LGB and has the early smoking initiation status corresponding to LGB status.

- $Y_{(0,M_0)}$, the potential outcome for adult smoking status when an individual is heterosexual and has the early smoking initiation status corresponding to heterosexual status.
- $Y_{(1,M_0)}$, the potential outcome for adult smoking status when an individual is LGB but has the early smoking initiation status corresponding to heterosexual status.
- $Y_{(0,M_1)}$, the potential outcome for adult smoking status when an individual is heterosexual but has the early smoking initiation status corresponding to LGB status.

As discussed previously, the latter two potential outcomes, $Y_{(1,M_0)}$ and $Y_{(0,M_1)}$, are never observed for any individual, yet allow us to more precisely define causal estimands for direct and indirect effects. We begin by defining the total effect (TE) of A on Y in the case of a binary exposure ($a = 1$ and $a' = 0$ or $a = 0$ and $a' = 1$):

$$TE = E\left(Y_{(a,M_a)} - Y_{(a',M_{a'})}\right) = E\left(Y_a - Y_{a'}\right) \tag{1}$$

where the expectation is over individuals.

The natural direct effect (NDE) and natural indirect effect (NIE), which sum to produce the total effect, are defined as follows:

$$\begin{aligned} NDE_{a'} &= E\left(\color{red}{Y_{(a,M_{a'})}}\color{black} - Y_{(a',M_{a'})}\right) & (2)\\ NIE_a &= E\left(Y_{(a,M_a)} - \color{red}{Y_{(a,M_{a'})}}\color{black}\right) & (3)\end{aligned}$$

Note that the NDE and NIE definitions rely on hypothetical (unobservable) potential outcomes, denoted in red and often referred to as cross-world counterfactuals or cross-world potential outcomes. The subscripts for NDE denote the condition to which the mediator is held constant, whereas the subscripts for NIE denote the condition to which the exposure is held constant. Each decomposition includes an NIE and an NDE corresponding to opposite subscripts.

As shown below, the NDE and NIE sum to the TE. Consider the following decomposition of TE in the case of a binary exposure for $a = 1$ and $a' = 0$, obtained by adding and subtracting $\color{red}{E\left(Y_{(1,M_0)}\right)}$:

$$\begin{aligned} \overbrace{E\left(Y_1 - Y_0\right)}^{\text{total effect}} &= E\left(Y_{(1,M_1)} - Y_{(0,M_0)}\right)\\ &= \overbrace{E\left(Y_{(1,M_1)} - Y_{(1,M_0)}\right)}^{\text{natural indirect effect}} + \overbrace{E\left(Y_{(1,M_0)} - Y_{(0,M_0)}\right)}^{\text{natural direct effect}}\\ &= NIE_1 + NDE_0 & (4)\end{aligned}$$

In the context of our motivating example, the NDE_0 term, $E\left(Y_{(1,M_0)} - Y_{(0,M_0)}\right)$, compares adult smoking status corresponding to LGB versus heterosexual status, holding early smoking initiation status to the value that would be obtained if heterosexual. The NDE_0 will be non-null only if LGB status has an effect on adult smoking status when early smoking initiation status is held fixed – namely, if LGB status has a **direct** effect on the outcome, not through the mediator.

The NIE_1 term $E\left(Y_{(1,M_1)} - Y_{(1,M_0)}\right)$ compares adult smoking status under the early smoking initiation status that would arise with and without the exposure condition (i.e., LGB status), for those in the exposure group (i.e., LGB women). The NIE_1 will be non-null only if LGB status has an **indirect** effect on adult smoking status via early smoking initiation among LGB women.

The previous TE decomposition comprised of NDE_0 and NIE_1 is obtained by adding and subtracting the term $\color{red}{E\left(Y_{(1,M_0)}\right)}$. We can similarly define an alternative TE decomposition comprised of NDE_1 and NIE_0, by adding and subtracting $\color{red}{E\left(Y_{(0,M_1)}\right)}$ as follows:

$$\begin{aligned}
\overbrace{E\left(Y_1 - Y_0\right)}^{\text{total effect}} &= E\left(Y_{(1,M_1)} - Y_{(0,M_0)}\right)\\
&= \overbrace{E\left(Y_{(1,M_1)} - Y_{(0,M_1)}\right)}^{\text{natural direct effect}} + \overbrace{E\left(Y_{(0,M_1)} - Y_{(0,M_0)}\right)}^{\text{natural indirect effect}}\\
&= NDE_1 + NIE_0
\end{aligned}\tag{5}$$

The `twangMediation` package provides estimates of both direct effects, NDE_0 and NDE_1, as well as both indirect effects, NIE_0 and NIE_1. Generally, if the treatment variable is defined as an exposure of interest versus a comparison group then the NIE_1 will be the mediating effect of interest. If the treatment variable reflects two alternative exposures of interest then the NIE_1 and NIE_0 are likely both of interest. See Nguyen et al. (2020) for a discussion of the differences between the two decompositions and how to decide which decomposition is of interest. For our case study, the NIE_1 is primarily the mediating effect of interest.

3 Identification Assumptions

In order to identify the natural (in)direct effects, we must impose assumptions that link the potential outcomes to our actual observed data. The approach implemented in `twangMediation` assumes positivity, consistency, and sequential ignorability, detailed below.

First, the positivity assumption requires that all individuals have some positive probability of receiving each level of the exposure and each level of the mediator. If individuals do not have a positive probability of receiving a particular level of the exposure or mediator, it is best to remove them from the sample because a causal effect is not meaningful for those individuals.

Additionally, the consistency assumption states that the outcome observed for an individual is identical to (i.e., consistent with) the potential outcome that corresponds to their observed exposure value; similarly, their observed mediator value is the potential mediator value that corresponds to their observed exposure value. In our example, if an individual's sexual identity is LGB ($A = 1$), then their observed mediator value M equals M_1 and their observed outcome Y equals $Y_{(1,M_1)}$. Similarly, if an individual's sexual identity is heterosexual ($A = 0$), then their observed mediator value M equals M_0 and their observed outcome Y equals $Y_{(0,M_0)}$.

Finally, sequential ignorability refers to a set of assumptions regarding confounding. The nonparametric assumptions typically made for identification of NDE and NIE conditioning on pre-exposure variables X are the following:

1. No unobserved confounding of the effect of A on M
2. No unobserved confounding of the effect of A on Y
3. No unobserved confounding of the effect of M on Y
4. No confounder (observed or unobserved) of the effect of M on Y that is affected by A

If individuals are randomly assigned to levels of the exposure, then assumptions 1 and 2 should hold. However, assumptions 3 and 4 may not hold even when there is random assignment to the exposure. See VanderWeele (2015) for further discussion of these identifying assumptions.

4 Estimation

The basic idea is to obtain estimates of $E\left(Y_{(1,M_1)}\right)$, $E\left(Y_{(0,M_0)}\right)$, $E\left(Y_{(1,M_0)}\right)$, and $E\left(Y_{(0,M_1)}\right)$ which are then plugged into Equations 4 or 5 to obtain estimates of the natural indirect and direct effects. Hong (2010) first defined the following weights $w_{aa'}$ to estimate each potential outcome, $E\left(Y_{(a,M_{a'})}\right)$:

$$w_{aa'} = \frac{p(M = m|A = a', X = x)}{p(M = m|A = a, X = x)p(A = a|X = x)}. \tag{6}$$

(Note that $w_{aa'}$ is a function of X as well as a and a' but we omit X from the $w_{aa'}$ notation for simplicity). Under the previously stated assumptions of consistency, positivity, and sequential ignorability (i.e., X strictly pre-exposure, or not affected by A), Huber (2014) used the following manipulation (i.e., Bayes Rule)

$$p(M = m|A = a, X = x) = \frac{p(A = a|M = m, X = x)p(M = m|X = x)}{p(A = a|X = x)}$$

to obtain an easier set of weights to estimate:

$$w_{aa'} = \frac{p(M = m|A = a', X = x)}{p(M = m|A = a, X = x)p(A = a|X = x)} = \overbrace{\frac{p(A = a'|M = m, X = x)}{p(A = a|M = m, X = x)}}^{\text{Odds Weight}} \overbrace{\frac{1}{p(A = a'|X = x)}}^{\text{IPW}} \tag{7}$$

These weights have been referred to as **cross-world weights** (Nguyen et al., 2021) as they are used to estimate the average cross-world potential outcomes (i.e., $E\left(Y_{(1,M_0)}\right)$ or $E\left(Y_{(0,M_1)}\right)$). In the denominator of Equation 7, note that $p(A = a|X = x)$ appears on the left hand side whereas $p(A = a'|X = x)$ appears on the right hand side; the change is the result of applying Bayes rule for the numerator and denominator of Equation 6. Following Nguyen et al. (2021), we will refer to the first term comprising the product on the right hand side of Equation 7 as an odds weight and the second term as an inverse probability weight (IPW). These terms are so named because the IPW is of the standard IPW form and the odds weight term is the usual form for estimating the average treatment effect among the treated/exposed (ATT), with the addition of conditioning on the mediator. In practice, the odds weight and IPW weight are calculated separately and then multiplied together to obtain the final cross-world weights.

As implemented in `twangMediation`, Generalized Boosted Modeling (GBM) is the default method used to estimate cross-world weights, whereas Huber (2014) used logistic or probit regression. As described below, `twangMediation` additionally provides the option to estimate the cross-world weights using logistic regression. Given that both TE decompositions (as shown in Equation 4 and Equation 5) may be of interest to the user, `twangMediation` estimates the required weights to estimate NDE_1, NDE_0, NIE_1 and NIE_0.

We begin with $E\left(Y_{(1,M_1)}\right)$ and $E\left(Y_{(0,M_0)}\right)$ – for these estimands, $a = a'$ in Equation 7. Consider the case of $a = a' = 1$.

$$w_{11} = \overbrace{\frac{p(A = 1|M = m, X = x)}{p(A = 1|M = m, X = x)}}^{\text{Odds Weight}} \overbrace{\frac{1}{p(A = 1|X = x)}}^{\text{IPW}} = \overbrace{1}^{\text{Odds Weight}} \overbrace{\frac{1}{p(A = 1|X = x)}}^{\text{IPW}} \tag{8}$$

As we can see, in this case, the odds weight term cancels out to become 1 and our final weight w_{11} is simply the standard IPW (i.e., IPW that would be used to balance non-randomized exposure groups in the absence of a mediator), estimated for the probability of $A = 1$. Similarly, when $a = a' = 0$, the odds weight term also cancels out to become 1 and our final weight w_{00} is the IPW, estimated for the probability of $A = 0$. In these cases where the final weight is equivalent to the corresponding IPW weight, we will refer to these weights as "total effect weights." We note that `twangMediation` does not estimate these total effect weights; rather, prior to using `twangMediation`, the user must estimate these weights (e.g., using a GBM propensity score model) and pass them to `twangMediation` (see Section 5.2). We emphasize that the user should check balance and diagnostics for the total effect weights prior to using `twangMediation`.

Next, we detail how `twangMediation` estimates the cross-world weights needed to obtain estimates of $E\left(Y_{(1,M_0)}\right)$ (for the decomposition in Equation 4) and $E\left(Y_{(0,M_1)}\right)$ (for the decomposition in Equation 5). Consider the case when $a = 0$ and $a' = 1$:

$$w_{01} = \overbrace{\frac{p(A=1|M=m,X=x)}{p(A=0|M=m,X=x)}}^{\text{Odds Weight}} \overbrace{\frac{1}{p(A=1|X=x)}}^{\text{IPW}} \tag{9}$$

To calculate the odds weight term, `twangMediation` calls the `ps` function in `twang` to estimate a propensity score model predicting membership in the $A = 1$ group based on the covariates X and mediator M. To calculate the IPW term, `twangMediation` calls the `ps` function in `twang` to estimate a propensity score model predicting membership in the $A = 1$ group based on the covariates X. The final cross-world weights w_{01} are calculated by multiplying the IPW with the respective odds weight term.

We note that although the IPW term in Equation 9 looks like the standard total effect weights provided by the user and used in Equation 8, `twangMediation` estimates this term in the context of Equation 9 to allow greater flexibility to the user. Specifically, this allows the user to use different covariates for the mediation analysis than for estimating the total effect weights, as might be appropriate if there are confounders related to the mediator and the outcome that do not confound the exposure and the outcome. Alternatively, if there is random assignment to the exposure, the user may wish to provide `twangMediation` with a vector of ones for the total effect weights but specify a non-null set of covariates X for the cross-world IPW. Additionally, this option allows the user to use different estimation methods for the total effect weights and the cross-world IPW.

Similarly, consider the case when $a = 1$ and $a' = 0$:

$$w_{10} = \overbrace{\frac{p(A=0|M=m,X=x)}{p(A=1|M=m,X=x)}}^{\text{Odds Weight}} \overbrace{\frac{1}{p(A=0|X=x)}}^{\text{IPW}} \tag{10}$$

To calculate the odds weight term of Equation 10, `twangMediation` calls the `ps` function in `twang` to estimate a propensity score model predicting membership in the $A = 0$ group based on the covariates X and mediator M. To calculate the IPW term, `twangMediation` calls the `ps` function in `twang` to estimate a propensity score model predicting membership in the $A = 0$ group based on the covariates X. The final cross-world weights w_{10} are calculated by multiplying the IPW with the respective odds weight term.

5 Using `twangMediation` for causal mediation

5.1 Overview of the `wgtmed` function

Below we detail the syntax for the `twangMediation` `wgtmed` function, which provides estimates of the total effect, natural indirect effects, and natural direct effects. The `wgtmed` function returns a `mediation` object. The `wgtmed` function is an extension of the `twang` `ps` function for estimating propensity score weights using GBM. As such, much of the syntax is similar between the `wgtmed` and `ps` functions. Please refer to the `twang` documentation for a comprehensive overview of the `ps` function.

Regarding data requirements, the `wgtmed` function works only with binary exposure variables. However, the mediator(s) may be defined as binary, ordinal, multinomial (categorical), or continuous variables. The ability to handle complex mediators is one of the advantages of specifying models for the exposure in the cross-world weights, rather than for the mediator as originally proposed by Hong (2010). The outcome may be defined as a binary or continuous variable. In our applied example, the exposure, mediator, and outcome are all binary variables. For analyses that include multiple mediators simultaneously, the mediators may be different variable types (e.g., a binary mediator and a continuous mediator). Missing data is allowed for covariates, but not the exposure, mediator, or outcome.

If you have not already done so, install `twangMediation` from CRAN by typing `install.packages("twangMediation")`. `twangMediation` relies on other R packages, especially `gbm`, `survey`, `twang`, and `lattice`. You may have to run `install.packages()` for these as well if they are not already installed. You will only need to do this step once. In the future, running `update.packages()` regularly will ensure that you have the latest versions of the packages, including bug fixes and new features. To start, load the `twangMediation` package. You may also need to load the `twang` package for estimating the total effect weights. You will have to do this step once for each new R session.

```
> library(twangMediation)
> library(twang)
```

The dataset for the motivating example described above is available with the package and is named `NSDUH_female`. The variable `lgb_flag` is the exposure, defined as 1 for LGB individuals and 0 for heterosexual individuals. The mediator, `cig15`, denotes early smoking initiation (prior to age 15), with 1=yes and 0=no. The outcome, `cigmon`, denotes adult smoking status (any past-month smoking), with 1=yes and 0=no. The remaining variables are potential confounders which will be used in estimating the weights.

```
> data(NSDUH_female)
```

The first analytic step is to estimate propensity score weights for the exposure (i.e., total effect weights). These are the usual inverse propensity weights which account for baseline differences across exposure groups. Note that these weights must be ATE weights rather than ATT weights. While these weights can be estimated in any manner, we demonstrate estimating these weights with GBM using the `twang` `ps` function. The first argument specifies a formula relating the exposure, `lgb_flag`, to the covariates that are used to generate the total effect weights. The code below generates an object `TEps` that contains the total effect weights that will be passed to the `wgtmed` function.

```
> TEps <- ps(formula = lgb_flag ~ age + race + educ + income + employ,
+             data=NSDUH_female, verbose=F, n.trees=6000, estimand="ATE", stop.method="ks.mean")
```

Next, we use the `wgtmed` function to obtain the mediation estimates of interest. The `wgtmed` function estimates the cross-world weights using GBM (although logistic regression may also be specified) and then estimates the total, natural direct, and natural indirect effects (using both the total effect weights and the cross-world weights). The `wgtmed` function returns a `mediation` object (here named `cig_med`). This estimation step is computationally intensive and can take a few minutes. We detail the required and optional arguments of this function below. Note that, while the default number of GBM trees is 10,000 in `wgtmed`, our sample code in this tutorial uses `ps_n.trees`=6000 to reduce computation time. In practice, when using a Windows machine, it may be necessary to increase the memory limit for R's working session using the `memory.limit()` function (e.g., `memory.limit(size = 32000)`).

```
> cig_med <- wgtmed(formula.med = cig15 ~ age + race + educ + income + employ,
+                  a_treatment="lgb_flag",
+                  y_outcome="cigmon",
+                  data=NSDUH_female,
+                  method="ps",
+                  total_effect_ps=TEps,
+                  total_effect_stop_rule="ks.mean",
+                  ps_version="gbm",
+                  ps_n.trees=6000,
+                  ps_interaction.depth=3,
+                  ps_shrinkage=0.01,
+                  ps_stop.method="ks.mean",
+                  ps_verbose=FALSE)
```

5.2 `wgtmed`: Required arguments

formula.med Specifies a formula relating the mediator, `cig15`, to the covariates that are used to estimate the cross-world weights. Note that a model predicting the mediator based on the specified covariates is never explicitly estimated; this formula notation is merely a convenient way to distinguish which variables are the mediator(s) versus the covariates. In our example, we use the same set of covariates to estimate both the total effect and the cross-world weights. However, if conceptually appropriate, the user can specify different covariates for the cross-world weight models (in `wgtmed`) and total effect models (estimated prior to running `wgtmed`). However, all variables used in the total effect model should appear in the model for the cross-world weights (but variables used in the cross-world weight model might not appear in the model for the total effect weights).

a_treatment Specifies the name of the treatment (exposure) variable, `lgb_flag`. The treatment variable must be defined as a 0/1 indicator. The variable name should be entered in quotes, as this argument expects a character string.

y_outcome Specifies the name of the outcome variable, `cigmon`. The variable name should be entered in quotes, as this argument expects a character string.

med_interact Specifies variables included in `formula.med` that equal interactions (or cross-products) of mediators and the other covariates. It should be `NULL` when there are no such interactions specified in `formula.med` and it should not include variables that equal interactions (or cross-products) of the other covariates. See the discussion of interactions later in the tutorial (Section 5.4) for further details on specifying interactions among variables in `wgtmed`.

data Specifies the name of the dataset.

method Specifies the method for estimating the cross-world weights. The default, `method = "ps"`, estimates the weights with GBM using the `ps` function in `twang`. If `method = "logistic"`, the weights are estimated using logistic regression, the approach originally proposed by Huber (2014). If `method = "crossval"`, the weights are estimated with GBM, but using cross-validation (rather than stopping rules) to choose the number of GBM iterations. For `method = "crossval"`, the number of cross-validation folds may be specified using the argument `ps_cv.folds`; the default is 10.

total_effect_ps; total_effect_weights The object that contains the total effect weights must be specified. If the twang ps function is used to estimate the total effect weights, the argument total_effect_ps is used to specify the ps object containing these weights; correspondingly, the total_effect_weights argument is left NULL. If total_effect_ps is specified, then the total_effect_stop_rule argument must also be included to specify which stopping rule (of those used in the ps call) should be used for the total effect weights. Alternatively, the user may specify a vector of total effect weights using the total_effect_weights argument; in this case, the total_effect_ps argument is left NULL. If total_effect_weights are provided, the user will get a warning that says "Reminder: Check that all confounders used to estimate supplied total effect weights are included as confounders in formula.med." We note that if the treatment condition was randomized, the vector of total effect weights may be set to 1 since the treatment groups would not be expected to differ with regard to covariates.

5.3 wgtmed: Optional arguments

ps_stop.method This argument allows the user to specify one or more stopping rules used to select the optimal number of GBM iterations for estimating the cross-world weights. The stopping rules are all metrics that quantify balance (or equivalence) between treatment groups with respect to the covariates. The package includes four built-in ps_stop.method objects: es.mean, es.max, ks.mean, and ks.max. The default is c("ks.mean", "ks.max"). Please refer to the twang documentation for further details.

ps_n.trees, ps_interaction.depth, ps_shrinkage These are parameters for the GBMs that wgtmed fits and stores when estimating the cross-world weights. The argument ps_n.trees specifies the maximum number of GBM iterations; the default is 10000. The ps_shrinkage argument controls the amount of shrinkage used for smoothing in the GBM algorithm. This argument must be a numeric value between 0 and 1 (denoting the learning rate); the default is 0.01. Small values such as 0.005 or 0.001 yield smooth fits but require greater values of ps_n.trees to achieve adequate fits. Computational time increases inversely with small values of the ps_shrinkage argument. wgtmed will issue a warning if the estimated optimal number of iterations is too close to the maximum number of GBM iterations, as this indicates that balance may improve if more complex models are considered − the user should increase ps_n.trees or increase ps_shrinkage if this warning appears. The argument ps_interaction.depth controls the level of interactions allowed in the GBMs; the default is 3.

ps_n.keep A numeric variable indicating the algorithm should only consider 1 of every ps_n.keep iterations of the propensity score model and optimize balance over this set instead of all iterations. Default: 1.

ps_version Specifies whether GBM is implemented using the R package gbm or the R package xgboost; the default is gbm.

ps_verbose This argument controls the amount of information printed to the console and is set to FALSE by default.

sampw Allows the user to specify sampling weights and is set to NULL by default.

There are several other more advanced arguments that are directly passed to the ps function including ps_perm.test.iters, ps_bag.fraction, ps_minobsinnode, ps_ks.exact, and ps_n.grid that are described in the main twang tutorial. All these arguments are optional and have specified defaults.

5.4 Interactions between the treatment and the mediator

In some applications, it may be appropriate to allow the relationship between a covariate and mediator to depend on the values of treatment. In linear models, such heterogeneity would be captured by an interaction between the covariate and treatment indicator in the mediator model. When weighting, the interactions need to be in the models for cross-world weights. Similarly, models might also include interactions between two or more covariates. The GBM model is sufficiently flexible to capture interactions without requiring they be explicitly specified. If GBM is used for estimating the conditional probability models through either `method=ps` or `method=crossval`, then no interactions must be explicitly specified. However, they can be, if the user wants to ensure they are included in the model. Furthermore, if `method=logistic`, then interactions must be specified. Unfortunately, specifying interactions with `wgtmed` is complicated.

First, as discussed previously in Section 4, `wgtmed` calculates the cross-world weights by estimating the conditional probability of treatment given the covariates and mediator or mediators. This means that although the concern is interactions between covariates and treatment, interactions must be specified as interactions between the mediator and covariates, not treatment and the covariates. Hence, if an analyst believes a covariate `X` would interact with treatment in a model for an mediator, then an interaction between “X” and the mediator should be included `formula.med`.

The second complication to specifying interactions is that GBM does not accept explicitly specified interactions, i.e., users will receive an error if they include an interaction specified with a “:” as in “X:M”. Hence, interactions must be specified by creating a variable in the input dataset that equals the product of covariates or the product of covariates and one or more mediators. These cross-product variables must then be specified in `formula.med`. This includes interactions that include the mediator. This might be confusing because `formula.med` specifies the meditiator as a function of covariates. This does not matter because `formula.med` is just used to identify mediator variables and covariates or interaction variables. Every interaction must be specified on the right-hand side of the $\sim$ in `formula.med`.

The final complication arises because calculation of the cross-world weights requires estimates of both the probability treatment conditional on the mediator and the covariates ($p(A|M, X)$) and the probability of treatment conditional on the covariates only ($p(A|X)$), as described in Section 4. The probability of treatment conditional on the covariates only cannot include any of the cross-products between the mediator and the covariates. However, because interactions are specified as variables, which equal the cross-products, `wgtmed` cannot distinguish interactions from other covariates. In particular, it cannot distinguish cross-products that include the mediator from other covariates. This is a problem because these cross-products must be in the model for treatment given the covariates and the mediator but *not* included in the model for treatment given only the covariates. To solve this problem, users must list in the parameter `med_interact` any variables that equal cross-products involving one or more mediators. For example, suppose we expect there might be an interaction between treatment and age in the model for the mediator `cig15` in the smoking case study. We then need to create a cross-product between `cig15` and `age`, `agecig15 = age * cig15` and “agecig15” must be included on the right-hand side of $\sim$ in `formula.med`

```
formula.med = cig15 ~ age + race + educ + income + employ + agecig15
```

and “agecig15”must be specified in `med_interact` as a string,

```
med_interact = ‘agecig15’                    .
```

We may also think that there may be an interaction between race and education of interest. Because this interaction does not include the mediator, we do not need to specify it in `med_interact`. We need to first create it as the cross-product between `race` and `educ`, `raceeduc = race * educ` and add it to the right-hand side of $\sim$ in `formula.med`

```
formula.med = cig15 ~ age + race + educ + income + employ + agecig15 + raceeduc.
```

All these complications can be avoided by using GBM and allowing it to identify interactions to include in the model, which might not include those being considered by the analyst.

6 Assessing balance diagnostics

6.1 Overview of balance in causal mediation context

Causal mediation analysis involves the comparison of groups with observed differences in their treatment (exposure) and mediator status. The key assumptions of causal mediation analysis is that conditional on observed covariates, those comparisons are unconfounded. As we detail below, analyst should assess whether the estimated casual mediation weights achieve adequate balance across treatment groups with respect to both the covariates and the mediator.

Checking the Covariate Distributions For the causal mediation weighting approach described in this tutorial, estimated weights must result in weighted distributions of the observed covariates that are balanced across treatment groups for each of the estimators (TE, NIE_1, NDE_0, NIE_1, and NDE_0). For example, NIE_1 is estimated by $\sum w_{i,11}Y_i / \sum w_{i,11} - \sum w_{i,10}Y_i / \sum w_{i,10}$ where summation is over the treatment group ($A = 1$). Hence, to avoid confounding the estimate of NIE_1 the distributions of the covariates for the treatment group weighted by w_{11} should match the distributions of the covariates for the treatment group weighted by w_{10}. Similar checks of covariate balance should be run for each of the other estimands.

Checking the Mediator Counterfactual Distributions In addition to covariate balance, one must also consider whether estimated weights have achieved adequate balance with regard to the mediator. Recall that NIE_1 is defined as $E\left(Y_{(1,M_1)} - Y_{(1,M_0)}\right)$. Weighting is supposed to weight the distribution of mediator M_1 values among the exposure group sample to match the distribution of the values of M_0 for the entire population to create the counterfactual distribution of $Y_{(1,M_0)}$. The distribution of mediator values for the comparison group sample, weighted by the total effect weights, estimates the distribution of M_0 for the total population. Hence, if the estimated cross-world weights, w_{10} are well-estimated then the distribution of the mediator in the exposure group weighted by the w_{10} weights should match the distribution of the mediator in the control group weighted by the total effect weights w_{00}. Likewise, estimating $E(Y_{(0,M_1)})$, NIE_0, and NDE_1 requires the counterfactual distribution of $Y_{(0,M_1)}$. It is estimated by weighting the mediator distribution in the control group to match the M_1 in the population. The distribution of M_1 in the population is estimated by the distribution of the mediator in the exposure group weighted by the total effect weight. Hence, if the estimated cross-world weights, w_{01} are well-estimated then the distribution of the mediator in the control group weighted by the w_{01} weights should match the distribution of the mediator in the exposure group weighted by the total effect weight w_{11}.

6.2 Balance tables using `bal.table.mediation` function

The analyst should perform balance diagnostic checks before interpreting the estimated mediation effects. The `twangMediation` function `bal.table.mediation` supports these balance checks. After estimating weights using `wgtmed`, one can use the `bal.table.mediation` function on the returned mediation object to obtain six balance tables (for each stopping rule): one unweighted balance table and one weighted balance table for the TE estimand (denoted `unw` and `ps`, respectively), and four weighted balance tables respectively corresponding to NIE_1, NDE_0, NIE_0, and NDE_1. The two tables labeled `TE` present covariate balance between treatment groups both in the unweighted data and using the total effect weights (w_{00}, w_{11}). The following 4 tables are similar to the total effect table, but check covariate balance using the weights for estimating NIE_1, NDE_0, NIE_0, and NDE_1, respectively. Weighted summaries are presented for each stopping rule selected in separate tables and labeled (e.g., *ks.mean*). These balance tables have the same format as covariate balance tables provided when using the `ps` command in `twang`.[1]

The `bal.table.mediation` function returns balance tables for TE, NIE_1, NDE_0, NIE_0, and NDE_1 comprised of the following columns:

tx.mn, ct.mn The mean for each covariate in the treatment (exposure) group, `tx.mn`, and control (comparison) group, `ct.mn`.

tx.sd, ct.sd The standard deviation for each covariate in the treatment group, `tx.sd`, and control group, `ct.sd`.

std.eff.sz The standardized mean difference is defined as the treatment group mean minus the control group mean divided by the control group standard deviation for the decomposition in Equation 4 and the treatment group standard deviation for the decomposition in Equation 5. If the standard deviation is very small, the resulting standardized mean difference will be very large; for readability, we set all standardized mean differences larger than 500 to `NA` (missing values).

stat, p Depending on whether the covariate is continuous or categorical, `stat` is a t-statistic or a χ^2 statistic corresponding to a statistical test of means across treatment groups. `p` is the associated p-value.

ks The Kolmogorov-Smirnov test statistic (testing for differences in the covariate distribution across treatment groups).

The function `bal.table.mediation` also provides checks of the weighted mediator distributions with a comparison of the weighted means and a KS statistic comparing the weighted distribution functions. These comparison are in the last two tables produced by the function labeled "Mediator distribution check: check_counterfactual_nie_1" for checking the weights used to estimate $E(Y_{(1,M_0)})$, NIE_1, and NDE_0, and "Mediator distribution check: check_counterfactual_nie_0" for checking the weights used to estimate $E(Y_{(0,M_1)})$, NIE_0 and NDE_1.

The `Mediator Distribution Check` tables are comprised of the following columns:

[1] When sampling weights are specified, i.e., `sampw` is not NULL, then the statistics for the "unweighted" tables are calculated using the sampling weights and the statistics for "weighted" tables use a composite of the sampling weights and either the total effect weights or the cross-world weights depending on the weights being evaluated. Also, when sampling weights are specified, they are used in the calculation of the cross-world weights and composites of the sampling weights and the total effect or cross-world weights are used to estimate the total, direct, and indirect effects.

cntfact.mn Mean of the mediator under the counterfactual condition. For NIE_1, this is the estimate of the (counterfactual) mean of the mediator under the comparison (control) condition, $E(M_0)$, estimated from the exposure (treatment) group – the cross-world-weighted mean for the exposure (treatment) group. For NIE_0, this is the estimate of the (counterfactual) mean of the mediator under the exposure (treatment) condition, $E(M_1)$, estimated from the comparison (control) group – the cross-world-weighted mean for the comparison group.

target.mn Mean of the mediator under the observed condition. For NIE_1, this is the mean of the mediator under the treatment condition estimated from the treatment group – the total effects weighted mean for the treatment group. For NIE_0, this is the mean of the mediator under the control condition estimated from the control group – the total effects weighted mean for the control group.

cntfact.sd, target.sd The weighted estimates of the standard deviations of the mediator distributions under the counterfactual and target (i.e., observed) groups.

std.eff.sz Standardized mean difference, which is now calculated between the counterfactual and target (i.e., observed) groups.

stat, p, ks Similarly, `stat` and `ks` now refer to statistical tests across counterfactual and target (i.e., observed) groups.

In addition to the tabular results from the `bal.table.mediation` function, `twangMediation` provides 2 balance diagnostic graphs:

1. **Covariate Standardized Effect Size Plot**: To request this plot, add the argument `plot = "TRUE"` to `bal.table.mediation()`. This figure shows standardized effect sizes for each covariate using weights for each of the TE, NIE_1, NDE_0, NIE_0, and NDE_1 estimands (as reported in the balance tables) to allow users to visually assess covariate balance after weighting.

2. **Mediator Density Plot**: To request this plot, use the `plot` function applied to the mediation object from `wgtmed`. This figure provides a visual check on the match of the weighted mediator distributions for both NIE_1 and NIE_0 (as reported in the balance tables). If the mediator is binary, then the plot is a bar chart; if mediator is continuous, the plot is a density curve. The plot is interactive: users must hit the `return` key to advance from the NIE_1 plot to the NIE_0 plot. The analyst should review the plot(s) corresponding to the NIE estimate(s) of interest.

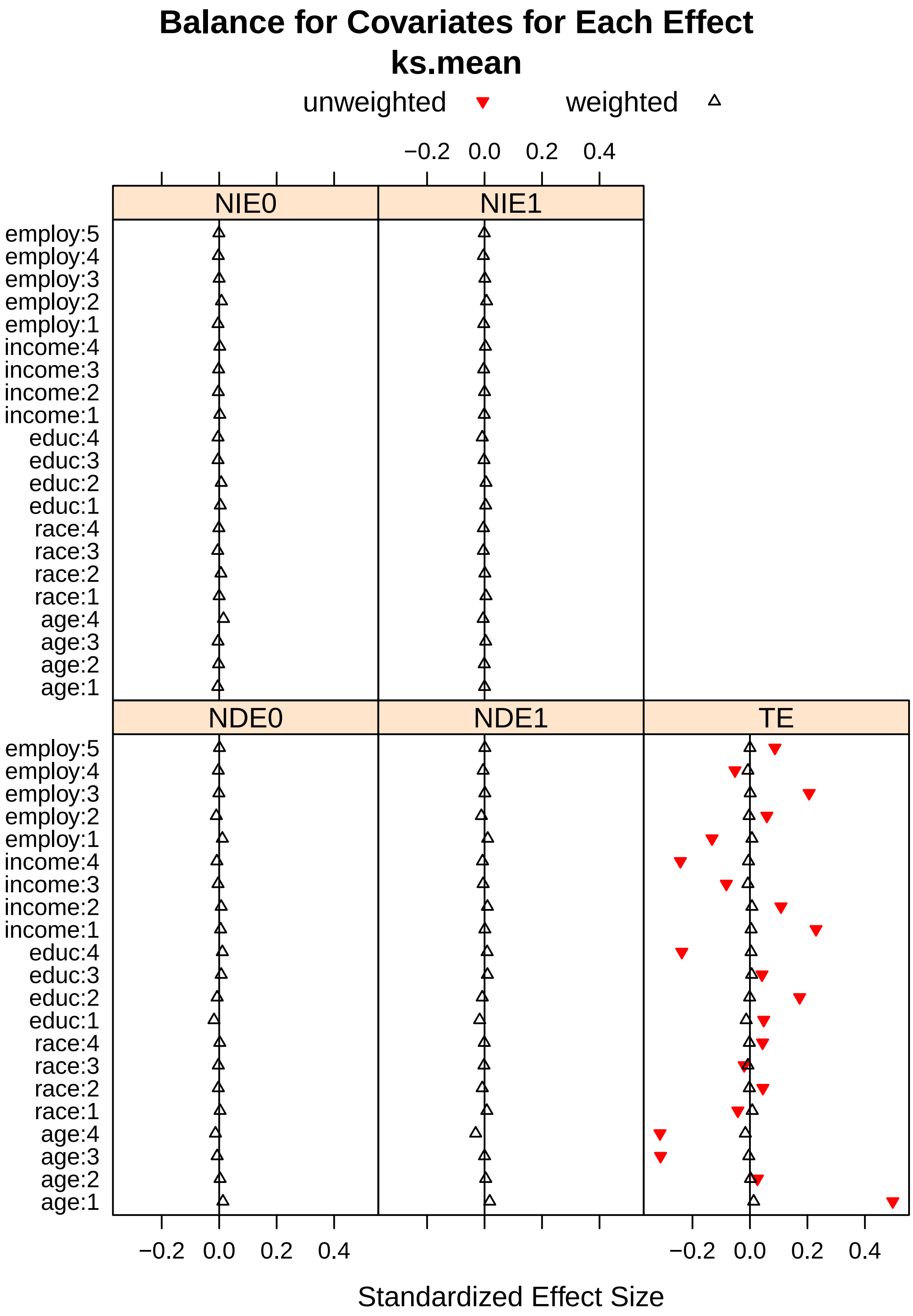
Balance diagnostics from our case study The balance table results for our applied example are shown below. We first examine the balance tables for total effects (TE). Prior to weighting, our two exposure groups (LGB women and heterosexual women) differed significantly with respect to all covariates (e.g., LGB women were younger and had lower household incomes than heterosexual women). After weighting, for all covariates, the absolute value of the `std.eff.sz` – known as absolute standardized mean difference (ASMD) – were well below 0.10. Next, since the NIE_1 and NDE_0 are the mediating effects of interest in our example, we examine the balance table for NIE_1, and NDE_0. Again, we see that weighting reduced the differences of all the covariates across exposure groups – all ASMDs were well below 0.10 after weighting.

Next we examine the Standardized Effect Size plot. For our example, as shown in the the total effects plot (labeled `TE`), the standard effect sizes for the covariates ranged from -0.31 to 0.49 prior to weighting, indicating significant differences between exposure groups. After weighting, the standard effect sizes are all near 0. In the plots for the natural indirect and direct effects both decompositions labeled NIE1, NDE0, NIE0, NDE1 the standard effect sizes are also close to 0 for all covariates, indicating weights removed any imbalances in the observed covariates for these estimands.

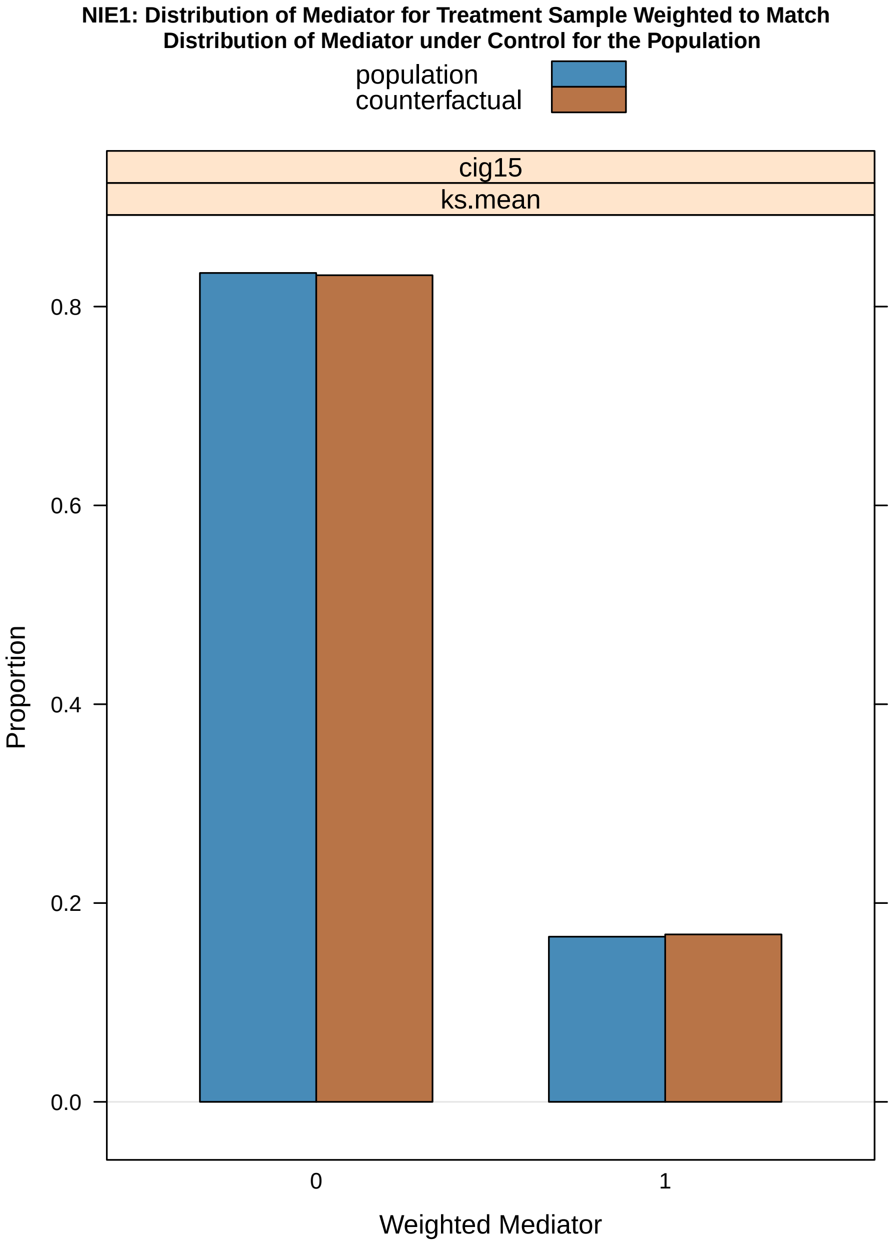
```
> bal.table.mediation(cig_med, plot = "TRUE")
------------------------------------------------------------------------------------------
Note: Balance for Covariates for Total Effects --
 "tx" treatment group weighted by w11 weights,
 "ct" control group weighted by w00 weights
------------------------------------------------------------------------------------------
$TE

$unw
         tx.mn tx.sd ct.mn ct.sd std.eff.sz    stat     p    ks
age:1    0.555 0.497 0.319 0.466      0.497 390.053 0.000 0.236
age:2    0.239 0.427 0.229 0.420      0.026      NA    NA 0.011
age:3    0.170 0.376 0.312 0.463     -0.311      NA    NA 0.142
age:4    0.035 0.185 0.140 0.347     -0.313      NA    NA 0.105
race:1   0.568 0.495 0.588 0.492     -0.042   5.673 0.001 0.021
race:2   0.149 0.356 0.133 0.340      0.045      NA    NA 0.015
race:3   0.172 0.377 0.180 0.384     -0.020      NA    NA 0.008
race:4   0.112 0.315 0.098 0.298      0.044      NA    NA 0.013
educ:1   0.121 0.326 0.106 0.308      0.048  81.281 0.000 0.015
educ:2   0.297 0.457 0.224 0.417      0.173      NA    NA 0.073
educ:3   0.383 0.486 0.363 0.481      0.042      NA    NA 0.020
educ:4   0.199 0.399 0.307 0.461     -0.237      NA    NA 0.108
income:1 0.295 0.456 0.201 0.401      0.230 116.806 0.000 0.094
income:2 0.351 0.477 0.302 0.459      0.108      NA    NA 0.050
income:3 0.125 0.331 0.154 0.361     -0.082      NA    NA 0.029
income:4 0.229 0.420 0.343 0.475     -0.242      NA    NA 0.114
employ:1 0.441 0.497 0.507 0.500     -0.132  56.751 0.000 0.066
employ:2 0.216 0.411 0.192 0.394      0.059      NA    NA 0.023
employ:3 0.098 0.297 0.050 0.219      0.206      NA    NA 0.047
employ:4 0.193 0.395 0.215 0.411     -0.052      NA    NA 0.021
employ:5 0.052 0.222 0.035 0.185      0.087      NA    NA 0.017

$ps
         tx.mn tx.sd ct.mn ct.sd std.eff.sz  stat     p    ks
age:1    0.349 0.477 0.343 0.475      0.013 0.209 0.839 0.006
age:2    0.230 0.421 0.230 0.421      0.002    NA    NA 0.001
age:3    0.296 0.456 0.298 0.457     -0.004    NA    NA 0.002
age:4    0.124 0.330 0.130 0.336     -0.016    NA    NA 0.005
race:1   0.590 0.492 0.586 0.492      0.008 0.060 0.981 0.004
race:2   0.134 0.341 0.135 0.342     -0.002    NA    NA 0.001
race:3   0.176 0.381 0.179 0.383     -0.007    NA    NA 0.003
race:4   0.099 0.299 0.100 0.300     -0.002    NA    NA 0.001
educ:1   0.104 0.305 0.108 0.310     -0.013 0.158 0.923 0.004
educ:2   0.231 0.421 0.232 0.422     -0.001    NA    NA 0.001
educ:3   0.367 0.482 0.365 0.481      0.006    NA    NA 0.003
educ:4   0.298 0.457 0.296 0.456      0.004    NA    NA 0.002
income:1 0.212 0.409 0.211 0.408      0.004 0.084 0.965 0.001
income:2 0.310 0.463 0.307 0.461      0.007    NA    NA 0.003
income:3 0.149 0.356 0.151 0.358     -0.007    NA    NA 0.002
income:4 0.329 0.470 0.331 0.471     -0.005    NA    NA 0.002
employ:1 0.504 0.500 0.501 0.500      0.007 0.056 0.991 0.004
employ:2 0.194 0.395 0.195 0.396     -0.003    NA    NA 0.001
employ:3 0.055 0.229 0.055 0.228      0.001    NA    NA 0.000
employ:4 0.210 0.407 0.213 0.409     -0.007    NA    NA 0.003
employ:5 0.037 0.189 0.037 0.189      0.000    NA    NA 0.000
------------------------------------------------------------------------------------------
Note: Balance for Covariates for NIE1 --
 "tx" treatment group weighted by w11 weights,
 "ct" treatment group weighted by w10 weights
------------------------------------------------------------------------------------------
$NIE1

$ks.mean
         tx.mn tx.sd ct.mn ct.sd std.eff.sz  stat     p    ks
age:1    0.349 0.477 0.349 0.477      0.000 0.005 0.997 0.000
age:2    0.230 0.421 0.231 0.421     -0.001    NA    NA 0.001
age:3    0.296 0.456 0.295 0.456      0.004    NA    NA 0.001
age:4    0.124 0.330 0.125 0.331     -0.005    NA    NA 0.001
race:1   0.590 0.492 0.588 0.492      0.005 0.015 0.997 0.002
race:2   0.134 0.341 0.134 0.341      0.001    NA    NA 0.000
race:3   0.176 0.381 0.178 0.382     -0.004    NA    NA 0.001
race:4   0.099 0.299 0.100 0.300     -0.004    NA    NA 0.001
educ:1   0.104 0.305 0.102 0.303      0.004 0.034 0.991 0.001
educ:2   0.231 0.421 0.229 0.420      0.005    NA    NA 0.002
educ:3   0.367 0.482 0.368 0.482     -0.002    NA    NA 0.001
educ:4   0.298 0.457 0.301 0.459     -0.008    NA    NA 0.003
income:1 0.212 0.409 0.213 0.409     -0.001 0.007 0.999 0.000
income:2 0.310 0.463 0.310 0.463      0.000    NA    NA 0.000
income:3 0.149 0.356 0.150 0.357     -0.003    NA    NA 0.001
income:4 0.329 0.470 0.327 0.469      0.003    NA    NA 0.001
employ:1 0.504 0.500 0.506 0.500     -0.003 0.029 0.997 0.002
employ:2 0.194 0.395 0.191 0.393      0.007    NA    NA 0.003
employ:3 0.055 0.229 0.055 0.228      0.001    NA    NA 0.000
employ:4 0.210 0.407 0.211 0.408     -0.004    NA    NA 0.002
employ:5 0.037 0.189 0.037 0.189     -0.001    NA    NA 0.000
------------------------------------------------------------------------------------------
Note: Balance for Covariates for NDE0 --
 "tx" treatment group weighted by w10 weights,
 "ct" control group weighted by w00 weights
------------------------------------------------------------------------------------------
$NDE0

$ks.mean
         tx.mn tx.sd ct.mn ct.sd std.eff.sz  stat     p    ks
age:1    0.349 0.477 0.343 0.475      0.013 0.177 0.868 0.006
age:2    0.231 0.421 0.230 0.421      0.003    NA    NA 0.001
age:3    0.295 0.456 0.298 0.457     -0.007    NA    NA 0.003
age:4    0.125 0.331 0.130 0.336     -0.013    NA    NA 0.004
race:1   0.588 0.492 0.586 0.492      0.003 0.018 0.997 0.001
race:2   0.134 0.341 0.135 0.342     -0.003    NA    NA 0.001
race:3   0.178 0.382 0.179 0.383     -0.003    NA    NA 0.001
race:4   0.100 0.300 0.100 0.300      0.002    NA    NA 0.001
educ:1   0.102 0.303 0.108 0.310     -0.018 0.357 0.782 0.006
educ:2   0.229 0.420 0.232 0.422     -0.007    NA    NA 0.003
educ:3   0.368 0.482 0.365 0.481      0.007    NA    NA 0.004
educ:4   0.301 0.459 0.296 0.456      0.011    NA    NA 0.005
income:1 0.213 0.409 0.211 0.408      0.005 0.086 0.964 0.002
income:2 0.310 0.463 0.307 0.461      0.007    NA    NA 0.003
income:3 0.150 0.357 0.151 0.358     -0.004    NA    NA 0.001
income:4 0.327 0.469 0.331 0.471     -0.008    NA    NA 0.004
employ:1 0.506 0.500 0.501 0.500      0.011 0.110 0.972 0.005
employ:2 0.191 0.393 0.195 0.396     -0.010    NA    NA 0.004
employ:3 0.055 0.228 0.055 0.228     -0.001    NA    NA 0.000
employ:4 0.211 0.408 0.213 0.409     -0.003    NA    NA 0.001
employ:5 0.037 0.189 0.037 0.189      0.001    NA    NA 0.000
------------------------------------------------------------------------------------------
Note: Balance for Covariates for NIE0 --
 "tx" control group weighted by w01 weights,
 "ct" control group weighted by w00 weights
------------------------------------------------------------------------------------------
$NIE0
```

```
$ks.mean
         tx.mn tx.sd ct.mn ct.sd std.eff.sz  stat     p    ks
age:1    0.341 0.474 0.343 0.475     -0.005 1.371 0.250 0.003
age:2    0.229 0.420 0.230 0.421     -0.002    NA    NA 0.001
age:3    0.296 0.456 0.298 0.457     -0.004    NA    NA 0.002
age:4    0.135 0.341 0.130 0.336      0.015    NA    NA 0.005
race:1   0.586 0.492 0.586 0.492      0.000 0.259 0.855 0.000
race:2   0.137 0.344 0.135 0.342      0.006    NA    NA 0.002
race:3   0.177 0.382 0.179 0.383     -0.005    NA    NA 0.002
race:4   0.099 0.299 0.100 0.300     -0.001    NA    NA 0.000
educ:1   0.109 0.311 0.108 0.310      0.004 0.411 0.745 0.001
educ:2   0.235 0.424 0.232 0.422      0.007    NA    NA 0.003
educ:3   0.363 0.481 0.365 0.481     -0.004    NA    NA 0.002
educ:4   0.294 0.456 0.296 0.456     -0.004    NA    NA 0.002
income:1 0.212 0.409 0.211 0.408      0.002 0.099 0.961 0.001
income:2 0.305 0.461 0.307 0.461     -0.003    NA    NA 0.001
income:3 0.151 0.358 0.151 0.358     -0.002    NA    NA 0.001
income:4 0.332 0.471 0.331 0.471      0.002    NA    NA 0.001
employ:1 0.499 0.500 0.501 0.500     -0.004 0.270 0.898 0.002
employ:2 0.198 0.398 0.195 0.396      0.008    NA    NA 0.003
employ:3 0.055 0.228 0.055 0.228      0.000    NA    NA 0.000
employ:4 0.212 0.408 0.213 0.409     -0.003    NA    NA 0.001
employ:5 0.037 0.188 0.037 0.189     -0.001    NA    NA 0.000
------------------------------------------------------------------------------------------
Note: Balance for Covariates for NDE1 --
 "tx" treatment group weighted by w11 weights,
 "ct" control group weighted by w01 weights
------------------------------------------------------------------------------------------
$NDE1

$ks.mean
         tx.mn tx.sd ct.mn ct.sd std.eff.sz  stat     p    ks
age:1    0.349 0.477 0.341 0.474      0.018 0.655 0.538 0.009
age:2    0.230 0.421 0.229 0.420      0.004    NA    NA 0.002
age:3    0.296 0.456 0.296 0.456      0.000    NA    NA 0.000
age:4    0.124 0.330 0.135 0.341     -0.031    NA    NA 0.010
race:1   0.590 0.492 0.586 0.492      0.008 0.068 0.977 0.004
race:2   0.134 0.341 0.137 0.344     -0.008    NA    NA 0.003
race:3   0.176 0.381 0.177 0.382     -0.002    NA    NA 0.001
race:4   0.099 0.299 0.099 0.299     -0.001    NA    NA 0.000
educ:1   0.104 0.305 0.109 0.311     -0.017 0.349 0.788 0.005
educ:2   0.231 0.421 0.235 0.424     -0.008    NA    NA 0.004
educ:3   0.367 0.482 0.363 0.481      0.010    NA    NA 0.005
educ:4   0.298 0.457 0.294 0.456      0.009    NA    NA 0.004
income:1 0.212 0.409 0.212 0.409      0.001 0.106 0.953 0.000
income:2 0.310 0.463 0.305 0.461      0.010    NA    NA 0.005
income:3 0.149 0.356 0.151 0.358     -0.005    NA    NA 0.002
income:4 0.329 0.470 0.332 0.471     -0.007    NA    NA 0.003
employ:1 0.504 0.500 0.499 0.500      0.011 0.129 0.964 0.006
employ:2 0.194 0.395 0.198 0.398     -0.011    NA    NA 0.004
employ:3 0.055 0.229 0.055 0.228      0.001    NA    NA 0.000
employ:4 0.210 0.407 0.212 0.408     -0.005    NA    NA 0.002
employ:5 0.037 0.189 0.037 0.188      0.001    NA    NA 0.000
------------------------------------------------------------------------------------------
Mediator Distribution Check: check_counterfactual_nie_1
------------------------------------------------------------------------------------------
        cntfact.mn cntfact.sd target.mn target.sd std.eff.sz   stat     p    ks
unw          0.269      0.444     0.167     0.373      0.267 14.234 0.000 0.102
ks.mean      0.168      0.374     0.166     0.372      0.006  0.333 0.739 0.002
------------------------------------------------------------------------------------------
Mediator Distribution Check: check_counterfactual_nie_0
------------------------------------------------------------------------------------------
```

```
        cntfact.mn cntfact.sd target.mn target.sd std.eff.sz    stat     p
unw          0.167      0.373     0.269     0.444     -0.267 -14.234 0.000
ks.mean      0.274      0.446     0.279     0.449     -0.014  -0.555 0.579
           ks
unw     0.102
ks.mean 0.005
------------------------------------------------------------------------------------------
```


Balance for Covariates for Each Effect
ks.mean
unweighted ▼
weighted △
NIE0
NIE1
NDE0
NDE1
TE
employ:5
employ:4
employ:3
employ:2
employ:1
income:4
income:3
income:2
income:1
educ:4
educ:3
educ:2
educ:1
race:4
race:3
race:2
race:1
age:4
age:3
age:2
age:1
−0.2
0.0
0.2
0.4
Standardized Effect Size

```
> plot(cig_med)
```



Finally, we examine the Mediator Density plot. As shown in the plot above, we see that weighted distributions for the population and the counterfactual distributions are well-matched in the context of NIE_1, indicating good weight performance. Based on these favorable diagnostic checks, we will proceed to our final effect estimates.

Obtaining the estimated weights Note that if the user wishes to obtain the estimated weights from `wgtmed` to construct their own plots or tables, they can be obtained as follows:

```
> w_00 <- attr(cig_med, 'w_00') #weight for estimating E[Y(0, M(0))]
> w_11 <- attr(cig_med, 'w_11') #weight for estimating E[Y(1, M(1))]
> w_10 <- attr(cig_med, 'w_10') #weight for estimating E[Y(1, M(0))]
> w_01 <- attr(cig_med, 'w_01') #weight for estimating E[Y(0, M(1))]
```

7 Interpreting the Effects: the `summary()` function

The `summary()` function applied to the mediation object from `wgtmed` provides a summary of all the important output including the effect estimates, covariate balance, effective sample size (ESS), and distribution checks for the mediator.

The ESS is reported because weighted means can have greater sampling variance than unweighted means from a sample of equal size. For example, the total effect and natural direct and indirect effects estimates equal differences of pairs of estimates of the four population means $E(Y_{(0,M_0)})$, $E(Y_{(1,M_1)})$, $E(Y_{(1,M_0)})$, and $E(Y_{(0,M_1)})$. Each population mean is estimated as a weighted mean. The means $E(Y_{(0,M_0)})$ and $E(Y_{(1,M_1)})$ use the appropriate total effect weights and the counterfactual means $E(Y_{(1,M_0)})$ and $E(Y_{(0,M_1)})$ use the corresponding cross-world weights. The variability of the weights will reduce the precision of the mean estimates and, subsequently, the estimated total, direct, and indirect effects. Large variability of the weights can also signal outliers where a small number of observations have very large weight relative to the average. The ESS is approximately the number of observations from a simple random sample that yields an estimate with sampling variation equal to the sampling variation obtained with the weighted comparison observations. It is an intuitive way to present the variability in the weights. Small values relative to the actual sample size indicate large variability in the weights, potential outliers, and possible low precision in the estimated mean and effect. This could signal the need to review data for the application. For each of the means:

$$ESS = \frac{\left(\sum_{i \in C} w_i\right)^2}{\sum_{i \in C} w_i^2} \tag{11}$$

where C is the set of indices for participants in the group used to estimate the mean, the exposure group for $E(Y_{(1,M_1)})$ and $E(Y_{(1,M_0)})$ or the comparison group for $E(Y_{(0,M_0)})$ and $E(Y_{(0,M_1)})$.[2]

The ESS for the four population means are presented in the second table in the output of the `summary` function. The output also includes the ESS for the odds weights and IPW weights used in calculating the cross-world weights. These ESSs are provided to help analysts diagnose the variability in the odds weight and IPW components to indicate the sources of variability in the cross-world weights and support model evaluation.

```
> summary(cig_med)
```

```
-----------------------------------------------------------------------------------------
95% Confidence Intervals for Effect Estimates: ks.mean_effects
-----------------------------------------------------------------------------------------
      effect std.err ci.min ci.max
TE     0.123   0.009  0.106  0.141
NDE_0  0.098   0.009  0.080  0.115
NIE_1  0.026   0.003  0.020  0.032
NDE_1  0.094   0.009  0.076  0.112
```

[2] The ESS is an accurate measure of the relative size of the variance of means when the weights are fixed or they are uncorrelated with outcomes. Otherwise the ESS is an underestimate (Little & Vartivarian, 2004). With propensity score weights, it is rare that weights are uncorrelated with outcomes. Hence, the ESS typically gives a lower bound, but it still serves as a useful measure for describing the variability of the weights and assessing the overall quality of a model, even if it provides a possibly conservative picture of the loss in precision due to weighting.

```
NIE_0  0.029   0.001  0.027  0.031
------------------------------------------------------------------------------------------
ESS for Total Effect and Cross-World Weights for estimating four population means used
to estimate the total effect and the natural direct and indirect effects
------------------------------------------------------------------------------------------
            E[Y(0, M(0))] E[Y(1, M(1))] E[Y(1, M(0))] E[Y(0, M(1))]
Sample Size      36163.00      4130.000      4130.000      36163.00
ks.mean          35981.46      2619.518      2519.793      32110.12
------------------------------------------------------------------------------------------
Balance Summary Tables: TE
Note: Balance for Covariates for Total Effects --
 "treat" treatment group weighted by w11 weights,
 "ctrl" control group weighted by w00 weights
------------------------------------------------------------------------------------------
    n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
unw    4130  36163  4130.000 36163.00  0.497   0.143  0.236   0.059
ps     4130  36163  2619.518 35981.46  0.016   0.006  0.006   0.002
------------------------------------------------------------------------------------------
Balance Summary Tables: NIE1
Note: Balance for Covariates for NIE1 --
 "treat" treatment group weighted by w11 weights,
 "ctrl" treatment group weighted by w10 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean    4130   4130  2619.518 2519.793  0.008   0.003  0.003   0.001
------------------------------------------------------------------------------------------
Balance Summary Tables: NDE0
Note: Balance for Covariates for NDE0 --
 "treat" treatment group weighted by w10 weights,
 "ctrl" control group weighted by w00 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean    4130  36163  2519.793 35981.46  0.018   0.007  0.006   0.003
------------------------------------------------------------------------------------------
Balance Summary Tables: NIE0
Note: Balance for Covariates for NIE0 --
 "treat" treatment group weighted by w01 weights,
 "ctrl" control group weighted by w00 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean   36163  36163  32110.12 35981.46  0.015   0.004  0.005   0.002
------------------------------------------------------------------------------------------
Balance Summary Tables: NDE1
Note: Balance for Covariates for NDE1 --
 "treat" treatment group weighted by w11 weights,
 "ctrl" control group weighted by w01 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean    4130  36163  2619.518 32110.12  0.031   0.008   0.01   0.003
------------------------------------------------------------------------------------------
```

The first table reports the total effect (TE), as well as the natural indirect and direct effects for both decompositions, NDE_0, NIE_1 and NDE_1, NIE_0, and their corresponding 95% confidence intervals. An NIE confidence interval that does not contain 0 indicates a statistically significant mediation effect at the 0.05 level. The next several tables are `Balance Summary Tables`, which offer a compact summary of sample sizes and balance measures for NIE_1, NDE_0, NIE_0, and NDE_1. The `Balance Summary Tables` are comprised of the following columns:

n.treat, n.ctrl The observed sample size in the exposure and comparison groups, respectively.

ess.treat, ess.ctrl The ESS after weighting for the exposure and comparison groups, respectively.

max.es, mean.es, max.ks, mean.ks Reports the maximum standardized mean difference, the mean standardized mean difference, the maximum KS statistic, and the mean KS statistic across all of the covariates, respectively. The last column, `iter`, gives the iteration number for each of the stop methods. This is not applicable to the unweighted model and thus, is given a value of `NA`.

We will now interpret the TE as well as the the decomposition of interest, NDE_0 and NIE_1, in the table labeled `95% Confidence Intervals for Effect Estimates` for our case study. Estimates from `wgtmed` are reported as marginal risk differences. The `TE` represents the total effect of LGB sexual identity on adult smoking status among women. The TE estimate of 0.123 represents a difference in magnitude of 12.3% in adult smoking rates between LGB and heterosexual women; statistical significance indicates that LGB women are significantly more likely than heterosexual women to be current smokers. The `NIE_1` is the natural indirect effect of early smoking initiation on adult smoking, holding LGB sexual identity ($A = 1$) constant. The NIE_1 estimate (0.026) is positive and statistically significant, indicating that early smoking initiation represents a significant pathway regarding adult smoking status (with early initiation accounting for approximately a 2.6% increase in magnitude in adult smoking rates). Examining the ratio of the `NIE_1` to the `TE` (0.026/0.123) indicates that approximately 21% of the total effect is through the mediator of early smoking initiation. The `NDE_0` is the natural direct effect of LGB status on smoking, holding early smoking initiation status constant to what it would be if a woman was heterosexual, $A = 0$. The NDE_0 is positive and statistically significant, indicating that LGB status is associated with smoking in adulthood, through mechanisms independent of early smoking initiation.

8 Estimating joint mediation effect of multiple mediators

Finally, we highlight that the `wgtmed` package can accept multiple mediators. When multiple mediators are included, the NIE and NDE estimands are calculated to reflect mediation jointly through all mediators (VanderWeele & Vansteelandt, 2014), rather than separate path-specific mediation effects (e.g., Daniel et al., 2015). The example below is an extension of our prior LGB disparities analysis. Now, our outcome is an indicator for whether an individual meets criteria for either alcohol or nicotine dependence `alc_cig_depend` and we consider 2 mediators: early smoking initiation `cig15` and early alcohol initiation `alc15`. To specify multiple mediators, include them on the left-hand side of the `formula.med` separated by "+".

```
>  TEps <- ps(lgb_flag ~ age + race + educ + income + employ,
+            data=NSDUH_female, verbose=F, n.trees=6000, n.keep=5, estimand="ATE")
>  cig_alc_med <- wgtmed(cig15 + alc15 ~ age + race + educ + income + employ,
+                  a_treatment="lgb_flag",
+                  y_outcome="alc_cig_depend",
+                  data=NSDUH_female,
+                  method="ps",
+                  total_effect_ps=TEps,
+                  total_effect_stop_rule="ks.mean",
+                  ps_version="gbm",
+                  ps_n.trees=6000,
+                  ps_n.keep = 5,
+                  ps_stop.method="ks.mean")

> summary(cig_alc_med)
------------------------------------------------------------------------------------------
95% Confidence Intervals for Effect Estimates: ks.mean_effects
------------------------------------------------------------------------------------------
      effect std.err ci.min ci.max
```

```
TE     0.084   0.008  0.068  0.099
NDE_0  0.059   0.008  0.044  0.074
NIE_1  0.025   0.003  0.018  0.032
NDE_1  0.056   0.008  0.041  0.072
NIE_0  0.027   0.001  0.025  0.030
------------------------------------------------------------------------------------------
ESS for Total Effect and Cross-World Weights for estimating four population means used
to estimate the total effect and the natural direct and indirect effects
------------------------------------------------------------------------------------------
            E[Y(0, M(0))] E[Y(1, M(1))] E[Y(1, M(0))] E[Y(0, M(1))]
Sample Size      36163.00      4130.000       4130.00      36163.00
ks.mean          35981.47      2619.652       2396.31      29702.21
------------------------------------------------------------------------------------------
Balance Summary Tables: TE
Note: Balance for Covariates for Total Effects --
 "treat" treatment group weighted by w11 weights,
 "ctrl" control group weighted by w00 weights
------------------------------------------------------------------------------------------
    n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
unw    4130  36163  4130.000 36163.00  0.497   0.143  0.236   0.059
ps     4130  36163  2619.652 35981.47  0.016   0.006  0.006   0.002
------------------------------------------------------------------------------------------
Balance Summary Tables: NIE1
Note: Balance for Covariates for NIE1 --
 "treat" treatment group weighted by w11 weights,
 "ctrl" treatment group weighted by w10 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean    4130   4130  2619.652  2396.31  0.011   0.005  0.004   0.002
------------------------------------------------------------------------------------------
Balance Summary Tables: NDE0
Note: Balance for Covariates for NDE0 --
 "treat" treatment group weighted by w10 weights,
 "ctrl" control group weighted by w00 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean    4130  36163   2396.31 35981.47  0.021   0.007  0.007   0.003
------------------------------------------------------------------------------------------
Balance Summary Tables: NIE0
Note: Balance for Covariates for NIE0 --
 "treat" treatment group weighted by w01 weights,
 "ctrl" control group weighted by w00 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean   36163  36163  29702.21 35981.47  0.023   0.006  0.008   0.002
------------------------------------------------------------------------------------------
Balance Summary Tables: NDE1
Note: Balance for Covariates for NDE1 --
 "treat" treatment group weighted by w11 weights,
 "ctrl" control group weighted by w01 weights
------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks mean.ks
ks.mean    4130  36163  2619.652 29702.21  0.039    0.01  0.013   0.004
------------------------------------------------------------------------------------------
```

As shown in the table above, the TE estimate of 0.084 represents a difference in magnitude of 8.4% in the rates of alcohol or nicotine dependence between LGB and heterosexual women, indicating a significant disparity. The NIE_1 estimate (0.025) is significant, indicating that early initiation of alcohol and smoking jointly represent a significant mediating pathway to adult dependence status among LGB women, with early initiation accounting for approximately a 2.5% increase in magnitude in adult dependence rates. Examining the ratio of NIE to TE, we conclude that early initiation accounts for approximately 30% of the adult disparity in alcohol or nicotine dependence. Additionally, the NDE_0 estimate is significant, indicating that LGB identity also has a significant effect on adult alcohol or nicotine dependence that is not attributed to early initiation of alcohol or smoking.

9 About this Tutorial

This tutorial was supported by funding from grant 1R01DA034065 from the National Institute on Drug Abuse. The overarching goal of the grant is to develop statistical methods and tools that will provide addiction health services researchers and others with the tools and training they need to study the effectiveness of treatments using observational data. The work is an extension of the Toolkit for Weighting and Analysis of Nonequivalent Groups, or TWANG, which contains a set of functions to support causal modeling of observational data through the estimation and evaluation of propensity score weights. The TWANG package was first developed in 2004 by RAND researchers for the R statistical computing language and environment and has since been expanded to include tools for SAS, Stata, and Shiny. For more information about TWANG and other causal tools being developed, see `www.rand.org/statistics/twang`.

RAND Social and Economic Well-Being is a division of the RAND Corporation that seeks to actively improve the health and social and economic well-being of populations and communities throughout the world. This research was conducted in the Social and Behavioral Policy Program within RAND Social and Economic Well-Being. The program focuses on such topics as risk factors and prevention programs, social safety net programs and other social supports, poverty, aging, disability, child and youth health and well-being, and quality of life, as well as other policy concerns that are influenced by social and behavioral actions and systems that affect well-being.

9.1 Acknowledgments

We would like to thank Trang Q. Nguyen, Emma Thomas, and Shu Xu for feedback on this tutorial and for beta testing the `twangMediation` package.

References

[1] Daniel, R. M., De Stavola, B. L., Cousens, S. N., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics, 71*, 1-15.

[2] Hong, G., Deutsch, J., & Hill, H. D. (2015). Ratio-of-Mediator-Probability Weighting for Causal Mediation Analysis in the Presence of Treatment-by-Mediator Interaction. *Journal of Educational and Behavioral Statistics, 40*(3), 307-340.

[3] Hong, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. *ASA Proceedings of the Joint Statistical Meetings*, pp. 2401–2415, American Statistical Association (Alexandria, VA)

[4] Huber, M. (2014). Identifying Causal Mechanisms (Primarily) Based on Inverse Probability Weighting. *Journal of Applied Econometrics, 29*(6), 920-943.

[5] Little, R. J., & Vartivarian, S. (2004). Does weighting for nonresponse increase the variance of survey means? *ASA Proceedings of the Joint Statistical Meetings*, 3897-3904 American Statistical Association (Alexandria, VA) `http://www.bepress.com/cgi/viewcontent.cgi?article=1034&context=umichbiostat`

[6] McCaffrey, D., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating adolescent substance abuse treatment. *Psychological Methods, 9*(4), 403-425.

[7] Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods, 26*(2), 255-271.

[8] Nguyen, T. Q., Ogburn, E. L., Sarker, E. B., Greifer, N., Schmid, I., Koning, I. M., & Stuart, E. A. (2021). Causal mediation analysis: From simple to more robust strategies for estimation of marginal natural (in)direct effects. `https://arxiv.org/abs/2102.06048v2`

[9] Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufman.

[10] Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3*(2), 143-155.

[11] Schuler, M. S., & Collins, R. L. (2019). Early alcohol and smoking initiation: A contributor to sexual minority disparities in adult use. *American Journal of Preventive Medicine, 57*(6), 808-817.

[12] VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction.* Oxford University Press.

[13] VanderWeele, T. J., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiol Methods, 2*(1), 95-115.