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

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

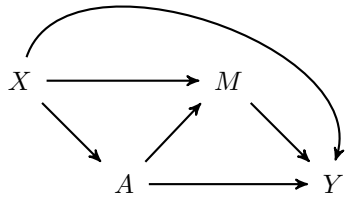
July 6, 2021

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 graph that illustrates a simple mediation model is shown below 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.



The **total effect** of A on Y includes two possible causal paths from A to Y: the path $A \rightarrow M \rightarrow Y$ is the **indirect effect** of A on Y through M and the path $A \rightarrow Y$ 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** mechanisms and causal effects 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.

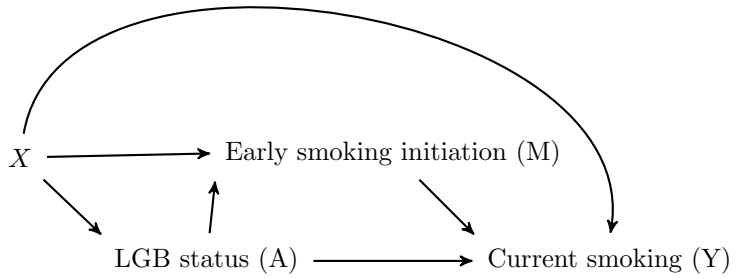
The mediator is an "intermediate" outcome of the exposure and itself has potential values. For each exposure level a there is a corresponding potential mediator value, denoted M_a. Also, there is a corresponding potential outcome that reflects the outcome value that would arise under the specific exposure 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. The following graph depicts our motivating example:

Figure 1: 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 $a' = 1$):

$$TE = Y_{i,(a,M_a)} - Y_{i,(a',M_{a'})} = Y_{i,a} - Y_{i,a'} \tag{1}$$

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

$$NDE_{a'} = Y_{i,(a,M_{a'})} - Y_{i,(a',M_{a'})} \tag{2}$$
$$NIE_{a} = Y_{i,(a,M_a)} - Y_{i,(a,M_{a'})} \tag{3}$$

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 $a' = 0$:

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

This decomposition is obtained by adding and subtracting $Y_{(1,M_0)}$, the potential outcome we would observe in a world where the exposure $A = 1$ and M is artificially manipulated to take the value it would naturally have under the condition $A = 0$.

In the context of our motivating example, the NDE_0 term, $Y_{(1,M_0)} - Y_{(0,M_0)}$, 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 individual 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 population version of this effect is $NDE_0 = E\left(Y_{(1,M_0)} - Y_{(0,M_0)}\right)$.

The NIE_1 term $Y_{(1,M_1)} - Y_{(1,M_0)}$ 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 individual 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 population version of this effect is $NIE_1 = E\left(Y_{(1,M_1)} - Y_{(1,M_0)}\right)$.

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

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

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 (in)direct

effects. Hong (2010) first defined the following weights w_i to estimate each potential outcome, $E\left(Y_{(a,M_{a'})}\right)$:

$$w_i = \frac{p(M_i = m|A_i = a', X_i = x)}{p(M_i = m|A_i = a, X_i = x)p(A_i = a|X_i = x)}. \tag{6}$$

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:

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

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 $E\left(Y_{(0,M_1)}\right)$). Note $p(A = a|X = x)$ in the denominator of the left hand side of Equation 7 compared to $p(A = a'|X = x)$ in the denominator of 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 in 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 may be of interest to the user, `twangMediation` estimates the required weights for both Equation 4 and Equation 5.

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 $a = a' = 1$.

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

As we can see, in this case, the odds weight term cancels out to become 1 and our final weight 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 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, they are estimated

previously (e.g., using a GBM propensity score model) and passed to `twangMediation` (see Section 5.1). We emphasize that the user 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 $a' = 1$.

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

To calculate the odds weight term, `twangMediation` calls the `ps` function in `twang` to estimate a propensity score model predicting membership in the *treated/exposed* 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 *treated/exposed* group based on the covariates X. The final cross-world weights 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 $a' = 0$.

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

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 *control/unexposed* 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 *control/unexposed* group based on the covariates X. The final cross-world weights are calculated by multiplying the IPW with the respective odds weight term.

5 Using `twangMediation` for causal mediation

Below we detail the syntax for the `wgtmed` function, which provides estimates of the total effect, natural indirect effects, and natural direct effects. The `twangMediation` `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 data 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 in a data frame named "w" 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, that we have named `cig_med`. This estimation step is computationally intensive and can take a few minutes. We set `ps_n.trees` to 6000 because we previously ran the function with 10000 and we know that the 6000 is sufficient for all the models. Thus, to reduce computation time in this tutorial, we reduced the number of trees from the default value of 10000. Note that, if 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)`). We detail the required and optional arguments of this function below.

```
> 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)
```

5.1 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 exposure variable, `lgb_flag`. The exposure 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.

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"`, then 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 or total_effect_weights The object that contains the total effect weights must be specified. The argument `total_effect_ps` is used to specify the `ps` object from estimating the total effect weights using the `twang ps` function, which contains the total effect 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 should be used for the total effect weights. Rather than specifying a `ps` object, the user may alternatively 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 to check that all confounders used for treatment (to obtain supplied

total effect weights) were included in confounders for the mediation model." We note that if the exposure condition was randomized, the vector of total effect weights may be set to 1 since the exposure groups would not be expected to differ with regard to covariates.

5.2 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 exposure groups with respect to the covariates. The `wgtmed` function selects the optimal number of GBM iterations to minimize the differences between exposure groups as measured by the rules of the given `ps_stop.method` object. 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 every ps_n.keep-th iteration 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, which we have not changed in this example.

5.3 Assessing balance diagnostics

The `wgtmed` function returns a `mediation` object. The analyst should perform several diagnostic checks before interpreting the estimated mediation effects.

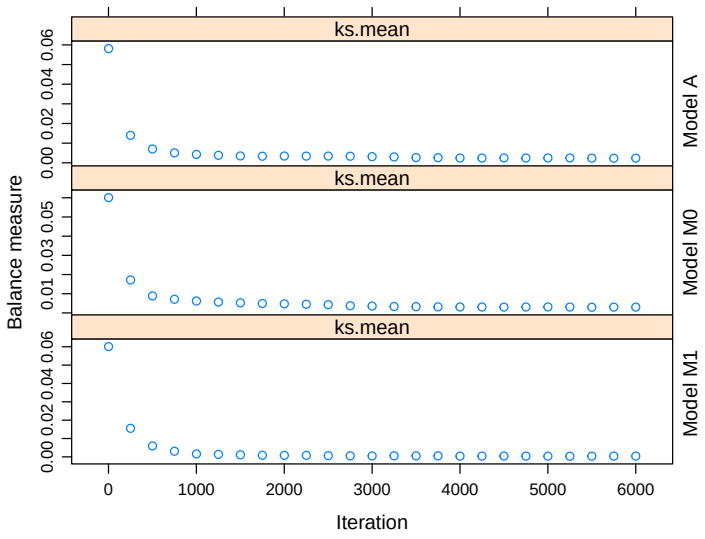
5.3.1 Diagnostic plots

GBM convergence plot: The first of these diagnostic plots assesses model convergence to make sure that the specified value of `ps_n.trees` allowed GBM to sufficiently explore complex models. The convergence plot is returned as the default of the `plot()` function or may be requested by specifying `plot="optimize"`. This convergence plot graphs the specified stopping criteria measure as a function of the number of iterations in the GBM algorithm, with higher iterations corresponding to more complex models. Note that this plot type is not available if `method=logistic` or `method=crossval`.

Three figures are displayed: (1) Model A, used for estimating the IPW term in Equation 7, (2) Model M0, used for estimating the odds weight term in Equation 7 for the NIE_1 and NDE_0 decomposition, and (3) Model M1, used for estimating the odds weight term in Equation 7 for the NIE_0 and NDE_1 decomposition. Adequate convergence should be achieved for all relevant models. If it appears that additional iterations would likely result in lower values of the balance statistic, `ps_n.trees` should be increased. However, after a point, additional complexity typically makes covariate balance worse. Note that the `model_subset` option can be used to display convergence plots for Model A, Model M0, or Model M1 individually.

If more than one stopping rule is specified in the `wgtmed` function, this figure will have multiple columns, corresponding to each stopping rule (unless the user specifies `subset` (e.g., subset=1 will only print plots for the first stopping rule)). This can be used to determine how comparable two or more stopping rules are: if the minima for multiple stopping rules under consideration are near one another, the results should not be sensitive to which stopping rule is used for the final analysis.

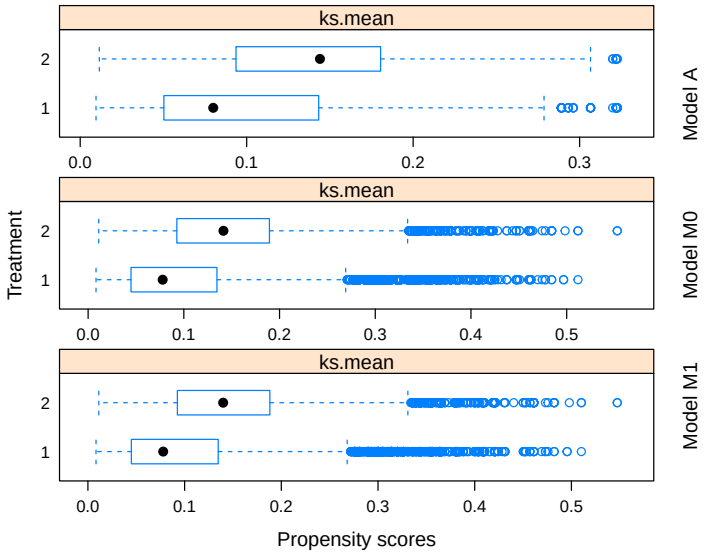
```
> plot(cig_med)
```



We are primarily interested in the NIE_1 and NDE_0 decomposition, (i.e., Equation 4), so we focus our interpretations on the plots labeled Model A and Model M0. As shown in the convergence plot above, we achieve good convergence for both Model A and Model M0, as the balance measures asymptotically approach 0 as the number of iterations increases.

Propensity score boxplots: The next plot produces boxplots illustrating the spread of the estimated propensity scores in the exposure and comparison groups for Model A, Model M0, and Model M1. If more than one stopping rule is specified in the `wgtmed` function, this figure will have multiple columns, corresponding to each stopping rule.

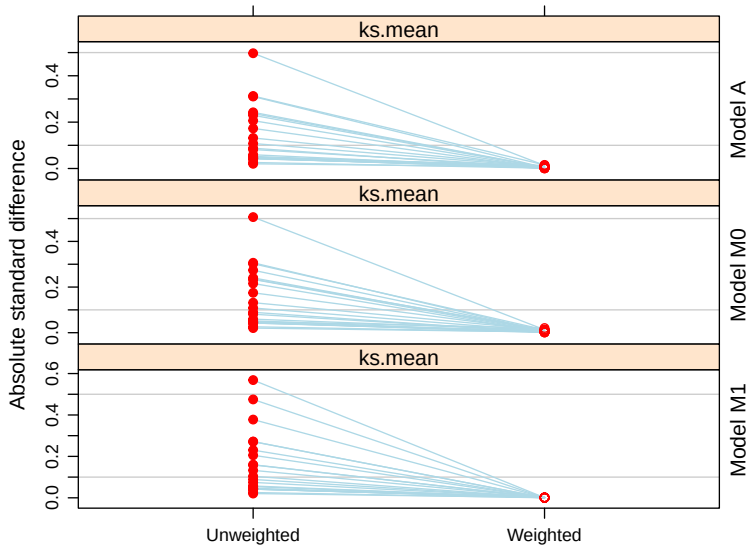
```
> plot(cig_med, plot = "boxplot")
```



As shown in the boxplot above, we see that we have adequate overlap in the propensity score values between the two exposure groups for Model A, Model M0, and Model M1.

Absolute standardized mean difference plots: The next plot illustrates the magnitude of the difference in covariate means across exposure groups, both before and after weighting. These magnitudes are reported using the absolute standardized mean difference (ASMD). In these plots, a blue line denotes a reduction in ASMD after weighting, whereas a red line denotes an increase. Closed red circles indicate a statistically significant difference in ASMD across groups. Ideally, the ASMD after weighting is less than 0.10 for all covariates.

```
> plot(cig_med, plot = "asmd")
```



As shown in the plot above, we see that we have sizable ASMD values in the unweighted data. However, weighting has notably decreased these covariate imbalances across exposure groups, as all ASMD values are well under 0.1 after weighting.

Mediator density plots: The `plot="density"` argument generates two figures assessing the distribution of the mediator variable in the context of NIE_1 and NIE_0, respectively. 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 see the next plot. The analyst should review the plot(s) corresponding to the NIE estimate(s) of interest.

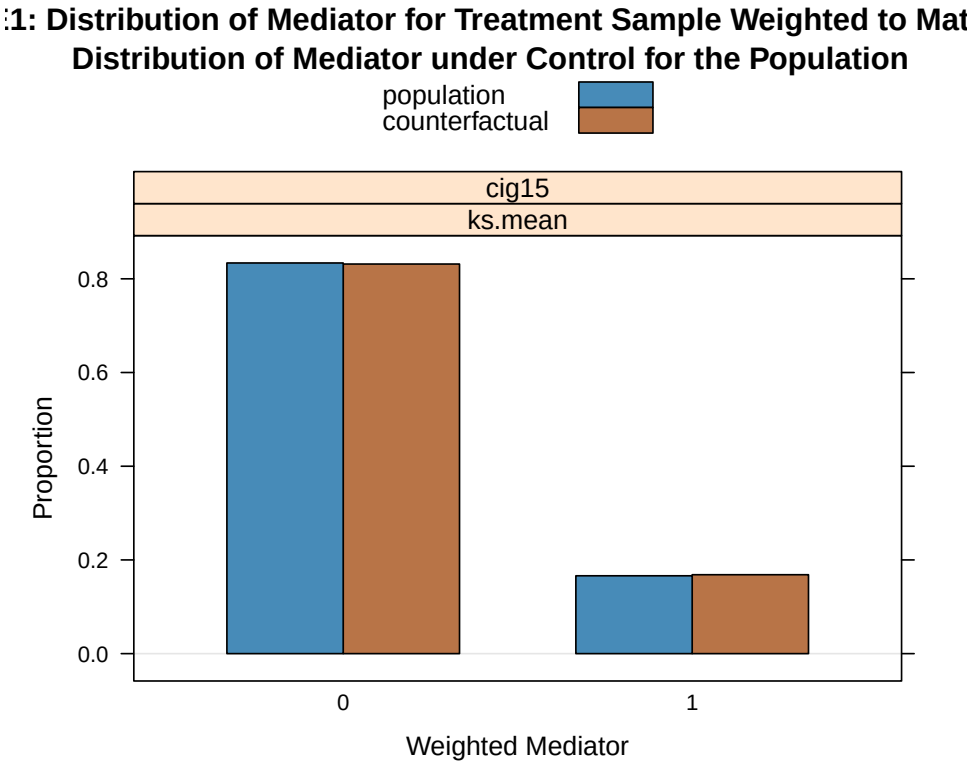
Recall that NIE_1 is defined as $E\left(Y_{(1,M_1)} - Y_{(1,M_0)}\right)$, hence it is defined among individuals in exposure group $A = 1$. Weighting is supposed to weight the distribution of mediator M_1 values among the exposure group $A = 1$ sample, to match the distribution of the values of M_0 for the entire population to create the counterfactual distribution of $Y_{(1,M_0)}$.

The distribution of mediator values for the comparison group, the $A = 0$ sample, weighted by the total effect weights estimates the distribution of M_0 for the total population. Hence, the first plot compares the distribution of mediator values for exposure group weighted by the cross-world weights (the counterfactual distribution, denoted "counterfactual" in the plot) to the distribution of mediator values for the comparison group sample weighted by the total effect weights (the population distribution of M_0, denoted "population" in the plot). Ideally, the weighted mediator variable distributions will be highly similar so that the counterfactual distribution and mean of $Y_{(1,M_0)}$ for the entire population is well-estimated by weighting the outcome of the exposure group by the cross-world weights and weighting has achieved its goal for estimating NIE_1 and NDE_0.

Similarly, the NIE_0 is defined as $E\left(Y_{(0,M_1)} - Y_{(0,M_0)}\right)$, hence it is defined among individuals in the comparison group $A = 0$. The second plot, which is for the NIE_0 and NDE_1 decomposition (i.e., Equation 5), compares the distribution of the mediator variable weighted by the cross-world weights among the observed comparison group (denoted "counterfactual" in the figure) to the distribution of the mediator variable weighted by the total effect weights in the exposed group (denoted "population" in the figure). Again, ideally, the weighted mediator

variable distributions will be highly similar so that the counterfactual distribution and mean of $Y_{(0,M_1)}$ for the entire population is well-estimated by weighting the outcome of the comparison group by the cross-world weights and weighting has achieved its goal for estimating NIE_0 and NDE_1.

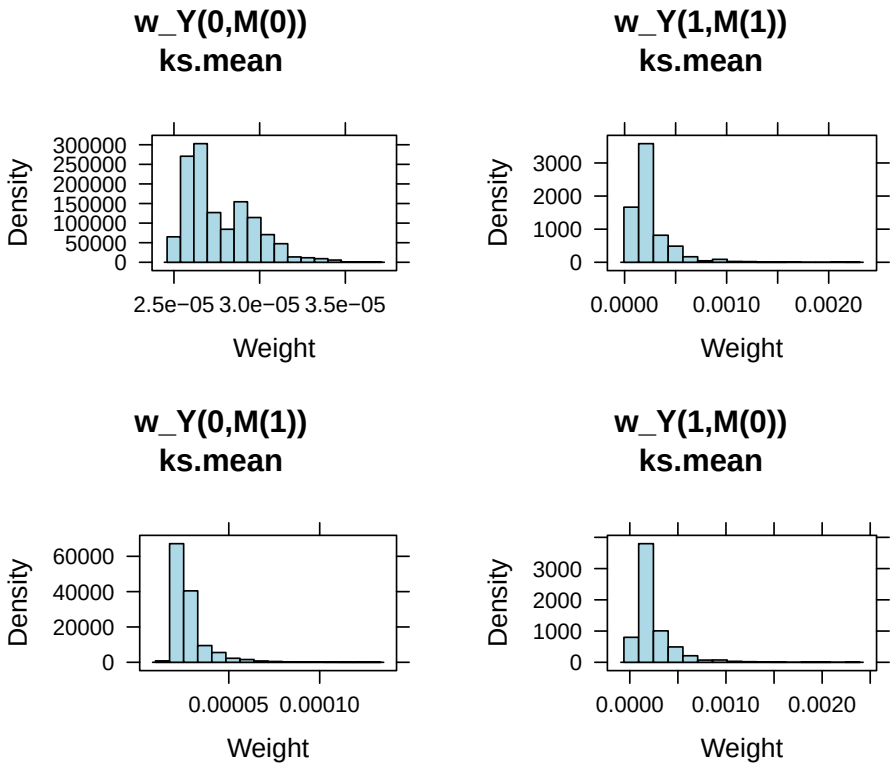
```
> plot(cig_med, plot = "density")
```



As shown in the density above, we see that weighted distributions for the population and the counterfactual distributions are well-matched in the context of NIE_1. In users' R console, pressing the return key would replace the plot with the density in the context of NIE_0.

Weight histograms: Finally, histograms of the standardized weights for each stopping rule can be obtained using the following code:

```
> plot(cig_med, plot = "weights")
```



The plot allows an assessment of the weights (e.g., identifying very large weights). If the user wishes to obtain the raw (not standardized) weights for further plotting, 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))]
```

5.3.2 Balance Tables

The function `bal.table.mediation()` applied to the mediation object returned by `wgtmed()` returns tables detailing covariate balance across exposure groups both before and after weighting. Three balance tables are presented, one for Model A, the model for the IPW term of the cross-world weights (denoted "balance_a"), and two for the odds weight (i.e., the first term) of the cross-world weights (denoted "balance_m0" and "balance_m1"). Ideally, weighting improves covariate balance across exposure groups and for the covariate and mediator balance after applying the cross-world weights.

The first table, `balance_a`, shows the balance between exposure groups both in the unweighted data and using the second term, the IPW, of the cross-world weights, and is relevant regardless of which total effect decomposition is of interest. This balance table is similar to the covariate balance table provided when using the `ps` command in `twang`.

Which of the second and third balance tables is relevant depends on the estimands (i.e., total effect decomposition) that the analyst wishes to use. We provide balance assessment for both

decompositions. If you wish to report the NIE_1 and NDE_0 estimands for the decomposition in Equation 4, then you will want to examine the table denoted `balance_m0`. If you wish to report the NIE_0 and NDE_1 estimands for the decomposition in Equation 5, then you will want to examine the table denoted `balance_m1`. For both tables, the weighted covariate summary statistics are calculated using weights of the form $\frac{p(A=a'|M=m,X=x)}{p(A=a|M=m,X=x)}$ (for NIE_1 and NDE_0) or $\frac{p(A=a|M=m,X=x)}{p(A=a'|M=m,X=x)}$ (for NIE_0 and NDE_1). Computationally, the `wtgmed` function estimates and optimizes GBM twice, once for the NIE_0 (using the original 0/1 coding of the exposure variable) and once for the NIE_1 (using reverse coding of the exposure variable), resulting in two balance tables labeled `balance_m1` and `balance_m0`, respectively.

The balance tables for Model A, Model M0, and Model M1 are comprised of the following columns:

tx.mn, ct.mn The mean for each covariate in the exposure group, `tx.mn`, and comparison group, `ct.mn`. The unweighted rows, denoted by the prefix `unw.`, show the unweighted means. Weighted summaries are presented for each stopping rule selected; the prefix corresponds to the specified stopping rule (e.g., `ks.mean.`).

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

std.eff.sz The standardized mean difference is defined as the exposure group mean minus the comparison group mean divided by the comparison group standard deviation for the decomposition in Equation 4 and the exposure 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 exposure groups. `p` is the associated p-value.

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

The balance table results for our applied example are shown below. We first examine the balance table for Model A. Prior to weighting, the two exposure groups differed significantly with respect to all covariates. After weighting, all ASMDs 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 Model M0. Again, we see significant differences between the two unweighted exposure groups with respect to the mediator variable as well as all covariates. Weighting reduced these differences across groups – all ASMDs were well below 0.10 after weighting.

```
> bal.table.mediation(cig_med)

********************************************************
Notes:
A. Model A estimates the probability of exposure given
the covariates specified in wgtmed. The results are used
by wgtmed to estimate E[Y(1,M(0))] and E[Y(0,M(1))].
They are not used to estimate the total effect.
B. Model M0 is used for NDE_0 and NIE_1 effects.
ct.sd is used for the denominator of std.eff.sz.
C. Model M1 is used for NDE_1 and NIE_0 effects.
tx.sd is used for the denominator of std.eff.sz.
See the bal.table.mediation help file for more information.
```

```
********************************************************

$balance_a
                tx.mn tx.sd ct.mn ct.sd std.eff.sz    stat  pval    ks
unw.age:1       0.555 0.497 0.319 0.466      0.497 390.053 0.000 0.236
unw.age:2       0.239 0.427 0.229 0.420      0.026      NA    NA 0.011
unw.age:3       0.170 0.376 0.312 0.463     -0.311      NA    NA 0.142
unw.age:4       0.035 0.185 0.140 0.347     -0.313      NA    NA 0.105
unw.race:1      0.568 0.495 0.588 0.492     -0.042   5.673 0.001 0.021
unw.race:2      0.149 0.356 0.133 0.340      0.045      NA    NA 0.015
unw.race:3      0.172 0.377 0.180 0.384     -0.020      NA    NA 0.008
unw.race:4      0.112 0.315 0.098 0.298      0.044      NA    NA 0.013
unw.educ:1      0.121 0.326 0.106 0.308      0.048  81.281 0.000 0.015
unw.educ:2      0.297 0.457 0.224 0.417      0.173      NA    NA 0.073
unw.educ:3      0.383 0.486 0.363 0.481      0.042      NA    NA 0.020
unw.educ:4      0.199 0.399 0.307 0.461     -0.237      NA    NA 0.108
unw.income:1    0.295 0.456 0.201 0.401      0.230 116.806 0.000 0.094
unw.income:2    0.351 0.477 0.302 0.459      0.108      NA    NA 0.050
unw.income:3    0.125 0.331 0.154 0.361     -0.082      NA    NA 0.029
unw.income:4    0.229 0.420 0.343 0.475     -0.242      NA    NA 0.114
unw.employ:1    0.441 0.497 0.507 0.500     -0.132  56.751 0.000 0.066
unw.employ:2    0.216 0.411 0.192 0.394      0.059      NA    NA 0.023
unw.employ:3    0.098 0.297 0.050 0.219      0.206      NA    NA 0.047
unw.employ:4    0.193 0.395 0.215 0.411     -0.052      NA    NA 0.021
unw.employ:5    0.052 0.222 0.035 0.185      0.087      NA    NA 0.017
ks.mean.age:1   0.349 0.477 0.343 0.475      0.013   0.209 0.839 0.006
ks.mean.age:2   0.230 0.421 0.230 0.421      0.002      NA    NA 0.001
ks.mean.age:3   0.296 0.456 0.298 0.457     -0.004      NA    NA 0.002
ks.mean.age:4   0.124 0.330 0.130 0.336     -0.016      NA    NA 0.005
ks.mean.race:1  0.590 0.492 0.586 0.492      0.008   0.060 0.981 0.004
ks.mean.race:2  0.134 0.341 0.135 0.342     -0.002      NA    NA 0.001
ks.mean.race:3  0.176 0.381 0.179 0.383     -0.007      NA    NA 0.003
ks.mean.race:4  0.099 0.299 0.100 0.300     -0.002      NA    NA 0.001
ks.mean.educ:1  0.104 0.305 0.108 0.310     -0.013   0.158 0.923 0.004
ks.mean.educ:2  0.231 0.421 0.232 0.422     -0.001      NA    NA 0.001
ks.mean.educ:3  0.367 0.482 0.365 0.481      0.006      NA    NA 0.003
ks.mean.educ:4  0.298 0.457 0.296 0.456      0.004      NA    NA 0.002
ks.mean.income:1 0.212 0.409 0.211 0.408     0.004   0.084 0.965 0.001
ks.mean.income:2 0.310 0.463 0.307 0.461     0.007      NA    NA 0.003
ks.mean.income:3 0.149 0.356 0.151 0.358    -0.007      NA    NA 0.002
ks.mean.income:4 0.329 0.470 0.331 0.471    -0.005      NA    NA 0.002
ks.mean.employ:1 0.504 0.500 0.501 0.500     0.007   0.056 0.991 0.004
ks.mean.employ:2 0.194 0.395 0.195 0.396    -0.003      NA    NA 0.001
ks.mean.employ:3 0.055 0.229 0.055 0.228     0.001      NA    NA 0.000
ks.mean.employ:4 0.210 0.407 0.213 0.409    -0.007      NA    NA 0.003
ks.mean.employ:5 0.037 0.189 0.037 0.189     0.000      NA    NA 0.000

$balance_m0
                tx.mn tx.sd ct.mn ct.sd std.eff.sz     stat  pval    ks
unw.cig15       0.269 0.444 0.167 0.373      0.274   14.234 0.000 0.102
unw.age:1       0.555 0.497 0.319 0.466      0.506 -390.053 0.000 0.236
unw.age:2       0.239 0.427 0.229 0.420      0.026       NA    NA 0.011
unw.age:3       0.170 0.376 0.312 0.463     -0.306       NA    NA 0.142
unw.age:4       0.035 0.185 0.140 0.347     -0.302       NA    NA 0.105
unw.race:1      0.568 0.495 0.588 0.492     -0.042   -5.673 0.001 0.021
unw.race:2      0.149 0.356 0.133 0.340      0.045       NA    NA 0.015
unw.race:3      0.172 0.377 0.180 0.384     -0.020       NA    NA 0.008
unw.race:4      0.112 0.315 0.098 0.298      0.044       NA    NA 0.013
unw.educ:1      0.121 0.326 0.106 0.308      0.048  -81.281 0.000 0.015
unw.educ:2      0.297 0.457 0.224 0.417      0.175       NA    NA 0.073
unw.educ:3      0.383 0.486 0.363 0.481      0.042       NA    NA 0.020
unw.educ:4      0.199 0.399 0.307 0.461     -0.235       NA    NA 0.108
unw.income:1    0.295 0.456 0.201 0.401      0.234 -116.806 0.000 0.094
```

```
unw.income:2     0.351 0.477 0.302 0.459      0.108       NA    NA 0.050
unw.income:3     0.125 0.331 0.154 0.361     -0.081       NA    NA 0.029
unw.income:4     0.229 0.420 0.343 0.475     -0.240       NA    NA 0.114
unw.employ:1     0.441 0.497 0.507 0.500     -0.132  -56.751 0.000 0.066
unw.employ:2     0.216 0.411 0.192 0.394      0.059       NA    NA 0.023
unw.employ:3     0.098 0.297 0.050 0.219      0.216       NA    NA 0.047
unw.employ:4     0.193 0.395 0.215 0.411     -0.052       NA    NA 0.021
unw.employ:5     0.052 0.222 0.035 0.185      0.089       NA    NA 0.017
ks.mean.cig15    0.170 0.376 0.167 0.373      0.007    0.389 0.697 0.003
ks.mean.age:1    0.325 0.468 0.319 0.466      0.013   -0.169 0.869 0.006
ks.mean.age:2    0.230 0.421 0.229 0.420      0.003       NA    NA 0.001
ks.mean.age:3    0.309 0.462 0.312 0.463     -0.006       NA    NA 0.003
ks.mean.age:4    0.136 0.343 0.140 0.347     -0.013       NA    NA 0.004
ks.mean.race:1   0.590 0.492 0.588 0.492      0.004   -0.021 0.996 0.002
ks.mean.race:2   0.132 0.339 0.133 0.340     -0.003       NA    NA 0.001
ks.mean.race:3   0.178 0.383 0.180 0.384     -0.003       NA    NA 0.001
ks.mean.race:4   0.099 0.299 0.098 0.298      0.002       NA    NA 0.000
ks.mean.educ:1   0.100 0.300 0.106 0.308     -0.021   -0.442 0.721 0.006
ks.mean.educ:2   0.221 0.415 0.224 0.417     -0.009       NA    NA 0.004
ks.mean.educ:3   0.367 0.482 0.363 0.481      0.008       NA    NA 0.004
ks.mean.educ:4   0.313 0.464 0.307 0.461      0.013       NA    NA 0.006
ks.mean.income:1 0.203 0.402 0.201 0.401      0.004   -0.078 0.968 0.002
ks.mean.income:2 0.305 0.460 0.302 0.459      0.007       NA    NA 0.003
ks.mean.income:3 0.153 0.360 0.154 0.361     -0.004       NA    NA 0.001
ks.mean.income:4 0.339 0.473 0.343 0.475     -0.008       NA    NA 0.004
ks.mean.employ:1 0.513 0.500 0.507 0.500      0.012   -0.135 0.958 0.006
ks.mean.employ:2 0.188 0.391 0.192 0.394     -0.012       NA    NA 0.005
ks.mean.employ:3 0.050 0.218 0.050 0.219     -0.002       NA    NA 0.000
ks.mean.employ:4 0.213 0.410 0.215 0.411     -0.003       NA    NA 0.001
ks.mean.employ:5 0.035 0.185 0.035 0.185      0.001       NA    NA 0.000

$balance_m1
                 tx.mn tx.sd ct.mn ct.sd std.eff.sz    stat  pval    ks
unw.cig15        0.269 0.444 0.167 0.373      0.230  14.234 0.000 0.102
unw.age:1        0.555 0.497 0.319 0.466      0.475 390.053 0.000 0.236
unw.age:2        0.239 0.427 0.229 0.420      0.026      NA    NA 0.011
unw.age:3        0.170 0.376 0.312 0.463     -0.378      NA    NA 0.142
unw.age:4        0.035 0.185 0.140 0.347     -0.569      NA    NA 0.105
unw.race:1       0.568 0.495 0.588 0.492     -0.042   5.673 0.001 0.021
unw.race:2       0.149 0.356 0.133 0.340      0.043      NA    NA 0.015
unw.race:3       0.172 0.377 0.180 0.384     -0.021      NA    NA 0.008
unw.race:4       0.112 0.315 0.098 0.298      0.042      NA    NA 0.013
unw.educ:1       0.121 0.326 0.106 0.308      0.046  81.281 0.000 0.015
unw.educ:2       0.297 0.457 0.224 0.417      0.160      NA    NA 0.073
unw.educ:3       0.383 0.486 0.363 0.481      0.042      NA    NA 0.020
unw.educ:4       0.199 0.399 0.307 0.461     -0.271      NA    NA 0.108
unw.income:1     0.295 0.456 0.201 0.401      0.205 116.806 0.000 0.094
unw.income:2     0.351 0.477 0.302 0.459      0.104      NA    NA 0.050
unw.income:3     0.125 0.331 0.154 0.361     -0.089      NA    NA 0.029
unw.income:4     0.229 0.420 0.343 0.475     -0.271      NA    NA 0.114
unw.employ:1     0.441 0.497 0.507 0.500     -0.133  56.751 0.000 0.066
unw.employ:2     0.216 0.411 0.192 0.394      0.057      NA    NA 0.023
unw.employ:3     0.098 0.297 0.050 0.219      0.159      NA    NA 0.047
unw.employ:4     0.193 0.395 0.215 0.411     -0.054      NA    NA 0.021
unw.employ:5     0.052 0.222 0.035 0.185      0.074      NA    NA 0.017
ks.mean.cig15    0.269 0.444 0.269 0.444      0.000   0.019 0.985 0.000
ks.mean.age:1    0.555 0.497 0.555 0.497      0.000   0.001 1.000 0.000
ks.mean.age:2    0.239 0.427 0.239 0.427      0.000      NA    NA 0.000
ks.mean.age:3    0.170 0.376 0.170 0.376      0.001      NA    NA 0.000
ks.mean.age:4    0.035 0.185 0.035 0.185     -0.001      NA    NA 0.000
ks.mean.race:1   0.568 0.495 0.569 0.495     -0.002   0.008 0.999 0.001
ks.mean.race:2   0.149 0.356 0.149 0.356      0.000      NA    NA 0.000
ks.mean.race:3   0.172 0.377 0.171 0.377      0.001      NA    NA 0.001
```

```
ks.mean.race:4   0.112 0.315 0.111 0.314      0.002      NA    NA 0.001
ks.mean.educ:1   0.121 0.326 0.121 0.326      0.001   0.004 1.000 0.000
ks.mean.educ:2   0.297 0.457 0.297 0.457      0.001      NA    NA 0.000
ks.mean.educ:3   0.383 0.486 0.384 0.486     -0.002      NA    NA 0.001
ks.mean.educ:4   0.199 0.399 0.199 0.399      0.000      NA    NA 0.000
ks.mean.income:1 0.295 0.456 0.295 0.456      0.000   0.001 1.000 0.000
ks.mean.income:2 0.351 0.477 0.351 0.477      0.000      NA    NA 0.000
ks.mean.income:3 0.125 0.331 0.125 0.331     -0.001      NA    NA 0.000
ks.mean.income:4 0.229 0.420 0.229 0.420      0.000      NA    NA 0.000
ks.mean.employ:1 0.441 0.497 0.442 0.497     -0.002   0.005 1.000 0.001
ks.mean.employ:2 0.216 0.411 0.216 0.411      0.000      NA    NA 0.000
ks.mean.employ:3 0.098 0.297 0.097 0.296      0.001      NA    NA 0.000
ks.mean.employ:4 0.193 0.395 0.193 0.395      0.001      NA    NA 0.000
ks.mean.employ:5 0.052 0.222 0.051 0.221      0.002      NA    NA 0.000

$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

$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
```

The final two tables are discussed in the next section.

5.4 Interpreting the Effects

The `summary()` function provides a summary of all the important output from `wgtmed` 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 $E(Y_{(0,M_1)})$.[1]

[1] The ESS is an accurate measure of the relative size of the variance of means when the weights are fixed or

The ESS for the four population means are presented in the table 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 diagnosis 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
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: model_a
Note: Model A is used for all effects: NDE_0, NDE_1, NIE_0, and NIE_1.
-------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks max.ks.p
unw        4130  36163  4130.000 36163.00  0.497   0.143  0.236       NA
ks.mean    4130  36163  2619.518 35981.46  0.016   0.006  0.006       NA
        mean.ks iter
unw       0.059   NA
ks.mean   0.002 5722
-------------------------------------------------------------------------------------------
Balance Summary Tables: model_m0
Note: Model M0 is used for NDE_0 and NIE_1 effects.
-------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks max.ks.p
unw        4130  36163  4130.000    36163  0.506   0.149  0.236       NA
ks.mean    4130  36163  2317.675    36163  0.021   0.007  0.006       NA
        mean.ks iter
unw       0.060   NA
ks.mean   0.003 5488
-------------------------------------------------------------------------------------------
Balance Summary Tables: model_m1
Note: Model M1 is used for NDE_1 and NIE_0 effects.
-------------------------------------------------------------------------------------------
        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks max.ks.p
unw        4130  36163      4130 36163.00  0.569   0.159  0.236       NA
ks.mean    4130  36163      4130 21529.22  0.002   0.001  0.001       NA
        mean.ks iter
unw        0.06   NA
ks.mean    0.00 5298
-------------------------------------------------------------------------------------------
Mediator Distribution Check: check_counterfactual_nie_1
-------------------------------------------------------------------------------------------
```

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.

```
        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
-----------------------------------------------------------------------------------------------
```

The first table reports the total effect (TE), as well as the natural (in)direct effects for both decompositions, NDE_0, NIE_1 and NDE_1, NIE_0, and their corresponding 95% confidence intervals in the table labeled `95% Confidence Intervals for Effect Estimates`. 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 Model A, Model M0, and Model M1. 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`.

The final two tables, labeled `Mediator Distribution Check`, have different columns. Which of these two tables is relevant again depends on the decomposition that the analyst is interested in. If the analyst is interested in the NIE_1 and NDE_0 estimands obtained from the decomposition in Equation 4, then the table labeled `check_counterfactual_nie_1` is relevant. If the analyst is interested in the NIE_0 and NDE_1 estimands obtained from the decomposition in Equation 5 then the table labeled `check_counterfactual_nie_0` is relevant. These tables show how well the cross-world weights achieve their goal of weighting the observed mediator (for one level of exposure, e.g., $A = 1$, M_1) to match the population of the potential mediator for the other exposure level (e.g., M_0), by checking that the cross-world-weighted mean and standard deviation of one sample (e.g., exposed) match the total-effect-weighted mean and standard deviation of the other sample (e.g., control).

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

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 condition ($E(M(0))$) estimated from the exposure group – the cross-world-weighted mean for the exposure group. For NIE_0, this is the estimate of the (counterfactual) mean of the mediator under the exposure condition ($E(M(1))$) estimated from the comparison 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 comparison condition estimated from the comparison group – the total effects weighted mean for the comparison group. For NIE_0, this is the mean of the mediator under the exposure condition estimated from the exposure group – the total effects weighted mean for the exposure 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.

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. The `TE` represents the total effect of LGB sexual identity on adult smoking status among women. As this is positive and statistically significant, LGB women are significantly more likely than heterosexual women to be current smokers. LGB women are estimated to be 12.4 percentage points more likely to report current smoking than heterosexual woman. 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. 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 is positive and statistically significant, indicating that indeed, LGB status is related to elevated smoking during adulthood through greater likelihood of early smoking initiation, which is positively associated with adult smoking. We note that, as expected, the `NDE_0` and `NIE_1` sum to the `TE` and roughly 21% of the total effect is through the mediator of early smoking initiation.

5.5 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 examining mediation effects of early smoking initiation on current (i.e., adult) smoking status. In the example below, we consider an additional mediator, early alcohol initiation `alc15`, in addition to early smoking initiation `cig15`. The outcome is an indicator for whether an individual meets criteria for either alcohol or nicotine dependence `alc_cig_depend`. 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: model_a
Note: Model A is used for all effects: NDE_0, NDE_1, NIE_0, and NIE_1.
-----------------------------------------------------------------------------------------------

        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks max.ks.p
unw        4130  36163  4130.000 36163.00  0.497   0.143  0.236       NA
ks.mean    4130  36163  2619.518 35981.46  0.016   0.006  0.006       NA
        mean.ks iter
unw       0.059   NA
ks.mean   0.002 5722
-----------------------------------------------------------------------------------------------
Balance Summary Tables: model_m0
Note: Model M0 is used for NDE_0 and NIE_1 effects.
-----------------------------------------------------------------------------------------------

        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks max.ks.p
unw        4130  36163  4130.000    36163  0.506   0.156  0.236       NA
ks.mean    4130  36163  2200.838    36163  0.021   0.008  0.007       NA
        mean.ks iter
unw       0.063   NA
ks.mean   0.003 6000
-----------------------------------------------------------------------------------------------
Balance Summary Tables: model_m1
Note: Model M1 is used for NDE_1 and NIE_0 effects.
-----------------------------------------------------------------------------------------------

        n.treat n.ctrl ess.treat ess.ctrl max.es mean.es max.ks max.ks.p
unw        4130  36163      4130 36163.00  0.569   0.163  0.236       NA
ks.mean    4130  36163      4130 20880.03  0.002   0.001  0.001       NA
        mean.ks iter
unw       0.063   NA
ks.mean   0.000 4208
-----------------------------------------------------------------------------------------------
Mediator Distribution Check: check_counterfactual_nie_1
-----------------------------------------------------------------------------------------------

             cntfact.mn cntfact.sd target.mn target.sd std.eff.sz   stat     p
unw.cig15         0.269      0.444     0.167     0.373      0.267 14.234 0.000
unw.alc15         0.256      0.436     0.144     0.351      0.309 15.892 0.000
ks.mean.cig15     0.165      0.371     0.166     0.372     -0.004 -0.233 0.816
ks.mean.alc15     0.147      0.354     0.144     0.351      0.008  0.513 0.608
                 ks
unw.cig15     0.102
unw.alc15     0.112
```

```
ks.mean.cig15 0.002
ks.mean.alc15 0.003
-----------------------------------------------------------------------------------------
Mediator Distribution Check: check_counterfactual_nie_0
-----------------------------------------------------------------------------------------
              cntfact.mn cntfact.sd target.mn target.sd std.eff.sz    stat
unw.cig15          0.167      0.373     0.269     0.444     -0.267 -14.234
unw.alc15          0.144      0.351     0.256     0.436     -0.309 -15.892
ks.mean.cig15      0.274      0.446     0.279     0.449     -0.015  -0.597
ks.mean.alc15      0.263      0.440     0.270     0.444     -0.020  -0.784
                   p    ks
unw.cig15     0.000 0.102
unw.alc15     0.000 0.112
ks.mean.cig15 0.551 0.006
ks.mean.alc15 0.433 0.007
-----------------------------------------------------------------------------------------
```

6 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.

6.1 Acknowledgments

We would like to thank Trang Q. Nguyen, Emma Thomas, Shu Xu, and Haoyu Zhou 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. (2020). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*.

[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.