mediator: an R package for conducting causal mediation analyses

Jordan Creed and Travis Gerke

Background: A mediation analysis examines the extent to which the causal relationship between an exposure and outcome operates through an intermediate variable, known as a mediator. A recently developed counterfactual framing of mediation analysis — referred to as *causal mediation analysis* — extends traditional methods by allowing exposure-mediator interactions and appropriately decomposing the total effect into direct and indirect effects.

Current software implementations for causal mediation analysis are available as SAS and SPSS macros by Valeri and VanderWeele. Within R, the package `mediation` (Tingley et al.) provides a subset of similar functions, but uses a different estimation approach and set of terminology which are less familiar to public health researchers.

We present the R package `mediator`, which provides point estimates and confidence intervals for statistics of interest in a causal mediation analysis. These include the controlled direct effect (CDE), natural direct and indirect effects (NDE and NIE), total effect (TE), and proportion mediated (PM). The package allows the user to specify whether an interaction between the exposure and mediator is assumed to exist. In addition to offering the first R implementation of causal mediation analysis as described by VanderWeele (2015), our implementation offers substantial performance enhancements over existing SAS/SPSS macros.

Usage: The package was developed using R v 3.6.1 and is currently available on GitHub. Installation and loading is accomplished with `devtools::install_github("gerkelab/mediator"); library(mediator)`. The `mediator` package allows for binary and continuous exposures, mediators and outcomes, as well as survival outcomes.

Use of the package is centered around the `mediator()` function. Minimum inputs for functionality include the analytic data set (`data =`), generlized linear model specifications for the outcome and mediator (`out.model =` and `med.model =`), and the treatment variable (`treat =`). Additional options include setting the exposure level (`a =`), the compared exposure level (`a_star =`), the level of the mediator (`m =`), and the number of bootstrap replications for calculating confidence intervals (`boot_rep =`). Fixed levels of the mediator are used for calculating CDE, hence, there are as many potential values for the CDE as there are levels of the mediator. By default the function calculates parametric confidence intervals using the Delta method, but user specification of nonzero bootstrap replicates automatically changes the method of confidence interval estimation to bootstrap.

The function returns a tibble with the CDE, NDE, NIE, TE, and PM, along with 95% confidence intervals. Returned effects are estimated at either the mean value (continuous) or the most common value (categorical) of covariates which are not the mediator or exposure variables.

Example: Using data on pancreatic cancer from The Cancer Genome Atlas (TCGA), we examine the effect of sex on overall survival and whether it is mediated through smoking status (never vs ever smoker). The proportion mediated is 0.09 and the direct effect of sex on pancreatic cancer survival is 0.80 and the indirect effect of sex on survival through smoking is 0.98, resulting in a combined total effect of 0.78. The controlled direct effect of male sex compared to female sex when forcing smoking status as never smokers is 0.79, while when setting smoking status as previous or current smoker the effect of male sex on survival status is 0.82.

Conclusion: The `mediator` R package provides an efficient mechanism for conducting causal mediation analyses and a useful tool in reproducible epidemiologic research.