SOCI 40258

Causal Mediation Analysis

Week 6: Estimating Interventional Effects

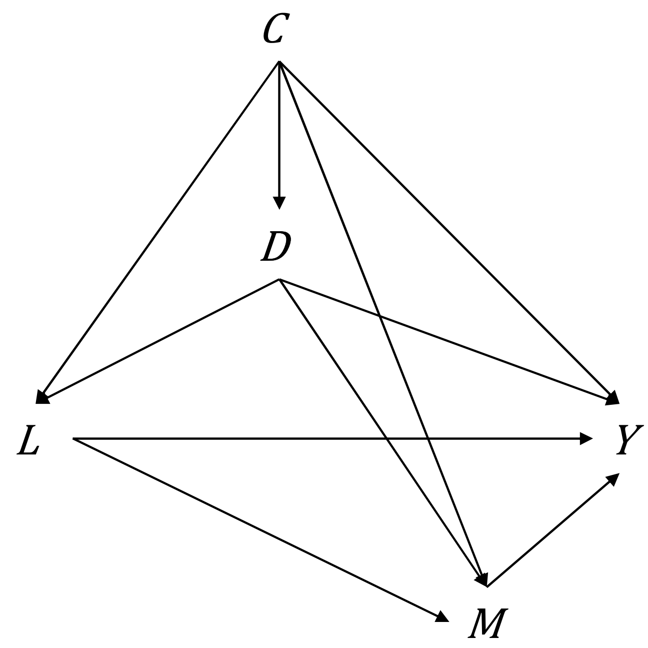
Outline

- Review of nonparametric identification and estimation
- Limitations of nonparametric estimation
- Parametric estimation with linear models
- Parametric estimation via simulation
- Parametric estimation via weighting

Models with multiple mediators

- In this model, the exposure D affects two mediators, L and M, which both affect the outcome Y, and L now affects M

- In this setting, we cannot use methods covered previously to analyze how M mediates the effect of D on Y
    - Natural effects cannot be nonparametrically identified in the presence of exposure-induced confounding



Models with multiple mediators

- The methods covered this week are appropriate for data arising from a causal process resembling the graphical model depicted previously
- My presentation of these methods is tailored for models that allow general patterns of both baseline and exposure-induced confounding
- These methods are also appropriate, however, for settings without any baseline confounding or without any exposure-induced confounding

The controlled direct effect

- The controlled direct effect:

$$CDE(d, d^*, m) = E\big(Y(d, m) - Y(d^*, m)\big)$$

- The $CDE(d, d^*, m)$ is the expected difference in the outcome if individuals had been exposed to d rather than d^* and if they had all experienced the same level of the mediator m

- It captures an effect of the exposure D on the outcome Y that persists after an intervention on the mediator M that sets, or controls, its value at the same level for everyone

Nonparametric identification

- Controlled direct effects can be nonparametrically identified if the following conditions are met:

Assumption CE.1: $Y(d,m) \perp D|C$

Assumption CE.2: $Y(d,m) \perp M|C,D,L$

Assumption CE.3: $P(d|c) > 0$ and $P(m|c,d,l) > 0$

Assumption CE.4: $Y = Y(D,M)$

Identification formula for CDE

- Under assumptions CE.1 to CE.4, the controlled direct effect can be equated with a function of observable data rather than the joint potential outcomes

- The nonparametric identification formula for the controlled direct effect can be expressed as follows:

$$CDE(d, d^*, m) = E\big(Y(d, m) - Y(d^*, m)\big)$$

$$= \textstyle\sum_c \sum_l [E(Y|c, d, l, m)P(l|c, d) - E(Y|c, d^*, l, m)P(l|c, d^*)]P(c)$$

$$= E_C\left(E_{L|c,d}\big(E(Y|C, d, L, m)\big) - E_{L|c,d^*}\big(E(Y|C, d^*, L, m)\big)\right)$$

Interventional effects decomposition

- An overall effect of the exposure on the outcome can be decomposed into interventional direct and indirect components as follows, defined in terms of randomized potential outcomes:

$$OE(d,d^*) = E\left(Y\big(d,\mathcal{M}(d|C)\big) - Y\big(d^*,\mathcal{M}(d^*|C)\big)\right)$$

$$= \underbrace{E\left(Y\big(d,\mathcal{M}(d^*|C)\big) - Y\big(d^*,\mathcal{M}(d^*|C)\big)\right)}_{\text{interventional direct effect}} + \underbrace{E\left(Y\big(d,\mathcal{M}(d|C)\big) - Y\big(d,\mathcal{M}(d^*|C)\big)\right)}_{\text{interventional indirect effect}}$$

The interventional direct effect

- The interventional direct effect:

$$IDE(d, d^*) = E\left(Y\big(d, \mathcal{M}(d^*|C)\big) - Y\big(d^*, \mathcal{M}(d^*|C)\big)\right)$$

- The $IDE(d, d^*)$ is the expected difference in the outcome if individuals had been exposed to d rather than d^* and if they had experienced the level of the mediator randomly drawn from its distribution under exposure d^*

- It captures an effect of the exposure D on the outcome Y that does not operate through its impact on the population distribution of the mediator M

The interventional indirect effect

- The interventional indirect effect:

$$IIE(d, d^*) = E\left(Y\big(d, \mathcal{M}(d|C)\big) - Y\big(d, \mathcal{M}(d^*|C)\big)\right)$$

- The $IIE(d, d^*)$ is the expected difference in the outcome if individuals had been exposed to d and then…
  - experienced a level of the mediator randomly drawn from its distribution under exposure d rather than from its distribution under exposure d^*

- It captures an effect of the exposure D on the outcome Y that arises from a shift in the population distribution of the mediator caused by a change in the exposure

The overall effect

- The overall effect:

$$OE(d, d^*) = E\left(Y\big(d, \mathcal{M}(d|C)\big) - Y\big(d^*, \mathcal{M}(d^*|C)\big)\right)$$

$$= IDE(d, d^*) + IIE(d, d^*)$$

- Similar to an average total effect, the $OE(d, d^*)$ it represents the expected difference in the outcome if…
  - individuals had been exposed to d rather than d^* and…
  - they had experienced a level of the mediator randomly selected from its distribution under exposure d as opposed to its distribution under exposure d^*
- It captures the effect of a change in the exposure and a corresponding shift in the population distribution of the mediator

Nonparametric identification

- Interventional effects can be nonparametrically identified if the following conditions are met:

Assumption IE.1: $Y(d,m) \perp D|C$

Assumption IE.2: $Y(d,m) \perp M|C,D,L$

Assumption IE.3: $M(d) \perp D|C$

Assumption IE.4: $P(d|c) > 0$ and $P(m|c,d,l) > 0$

Assumption IE.5: $Y = Y(D,M)$ and $M = M(D)$

Identification formula for IDE

- Under assumptions IE.1 to IE.5, the interventional direct effect can be equated with a function of observable data rather than the randomized potential outcomes
- The nonparametric identification formula for the interventional direct effect can be expressed as follows:

$$IDE(d,d^*) = E\left(Y\big(d,\mathcal{M}(d^*|C)\big) - Y\big(d^*,\mathcal{M}(d^*|C)\big)\right)$$

$$= \textstyle\sum_c \sum_m \sum_l [E(Y|c,d,l,m)P(l|c,d) - E(Y|c,d^*,l,m)P(l|c,d^*)]P(m|c,d^*)P(c)$$

$$= E_C\left(E_{M|C,d^*}\left(E_{L|C,d}\big(E(Y|C,d,L,M)\big) - E_{L|C,d^*}\big(E(Y|C,d^*,L,M)\big)\right)\right)$$

Identification formula for IIE

- Under the same set of assumptions, the interventional indirect effect can also be equated with a function of observable data rather than the randomized potential outcomes

- The nonparametric identification formula for the interventional indirect effect can be expressed as follows:

$$IIE(d, d^*) = E\left(Y\big(d, \mathcal{M}(d|C)\big) - Y\big(d, \mathcal{M}(d^*|C)\big)\right)$$

$$= \sum_c \sum_m \sum_l [P(m|c,d) - P(m|c,d^*)]\, E(Y|c,d,l,m) P(l|c,d) P(c)$$

$$= E_C\left(E_{M|C,d}\left(E_{L|C,d}\big(E(Y|C,d,M)\big)\right) - E_{M|C,d^*}\left(E_{L|C,d}\big(E(Y|C,d,M)\big)\right)\right)$$

Nonparametric estimation

- Nonparametric estimation involves plugging in sample analogs for the population quantities in the nonparametric identification formulae outlined previously

- However, this approach to estimation is often difficult, impractical, or even impossible to implement owing to:
  - sparsity
  - the curse of dimensionality
  - excessive sampling variability

Parametric estimation

- A parametric estimator is based on a parametric model or a set of parametric models (i.e., models that impose constraints on the joint distribution of the data)

- Parametric estimators can mitigate the problems of sparsity, dimensionality, and variability by sharing, borrowing, and filling in information that is otherwise not available from the sample data

  - With parametric estimation, the positivity condition required for nonparametric identification and estimation is supplanted by an alternative assumption about correct model specification

  - Parametric estimators are only consistent if their underlying models are correctly specified; otherwise, they are biased

Regression-with-residuals (RWR)

- Consider the following set of linear models:

$$E(L|c,d) = \lambda_0 + \lambda_1^T c^{\perp} + \lambda_2 d$$

$$E(M|c,d) = \beta_0 + \beta_1^T c^{\perp} + \beta_2 d$$

$$E(Y|c,d,l,m) = \gamma_0 + \gamma_1^T c^{\perp} + \gamma_2 d + m(\gamma_3 + \gamma_4 d) + \gamma_5^T l^{\perp},$$

where $c^{\perp} = c - \bar{C}$ and $l^{\perp} = l - \hat{E}(L|c,d)$

- Under these models, the interventional effects of interest are given by:

$$\left.\begin{aligned} IDE(d,d^*) &= \big(\gamma_2 + \gamma_4(\beta_0 + \beta_2 d^*)\big)(d - d^*) \\ IIE(d,d^*) &= \beta_2(\gamma_3 + \gamma_4 d)(d - d^*) \end{aligned}\right\} \text{notice anything familiar?}$$

$$CDE(d,d^*,m) = (\gamma_2 + \gamma_4 m)(d - d^*)$$

RWR with covariate interactions

- RWR easily accommodates two-way interactions of $c^{\perp} = c - \bar{C}$ and $l^{\perp} = l - \hat{E}(L|c,d)$ with d and m

- Under models with these interactions, the effects of interest are still given by the following parametric expressions:

$$\left.\begin{aligned} IDE(d,d^*) &= \big(\gamma_2 + \gamma_4(\beta_0 + \beta_2 d^*)\big)(d - d^*) \\ IIE(d,d^*) &= \beta_2(\gamma_3 + \gamma_4 d)(d - d^*) \\ CDE(d,d^*,m) &= (\gamma_2 + \gamma_4 m)(d - d^*) \end{aligned}\right\} \text{identical to the previous slide}$$

RWR estimation

- The RWR estimator can be implemented through the following series of steps:
  1. Center the baseline confounders around their sample means
  2. Residualize the exposure-induced confounders with respect to the observed past
  3. Fit a linear model for the mediator using the terms from step 1
  4. Fit a linear model for the outcome using the terms from steps 1 and 2
  5. Construct effect estimates from the model parameters in steps 3 and 4

RWR estimation

- Step 1: center the baseline confounders around the sample means
  - For each baseline confounder, compute $C^{\perp} = C - \bar{C}$, which centers these variables around their sample means
- Step 2: residualize the exposure-induced confounders
  - For each exposure-induced confounder, compute $L^{\perp} = L - \hat{E}(L|C, D)$ by fitting a least squares regression of L on C and D and then extracting the residuals

RWR estimation

- Step 3: fit a linear model for the mediator
  - Using $C^{\perp}$ from step 1, compute least squares estimates for the mediator model, which can be expressed as follows:

$$\hat{E}(M|C,D) = \hat{\beta}_0 + \hat{\beta}_1^T C^{\perp} + \hat{\beta}_2 D$$

- Step 4: fit a linear model for the outcome
  - Using $C^{\perp}$ and $L^{\perp}$ from steps 1 and 2, compute least squares estimates for the outcome model, which can be expressed as follows:

$$\hat{E}(Y|C,D,L,M) = \hat{\gamma}_0 + \hat{\gamma}_1^T C^{\perp} + \hat{\gamma}_2 D + M(\hat{\gamma}_3 + \hat{\gamma}_4 D) + \hat{\gamma}_5^T L^{\perp}$$

RWR estimation

- Step 5: construct effect estimates from the model parameters

$$\widehat{IDE}(d, d^*) = \left(\hat{\gamma}_2 + \hat{\gamma}_4\left(\hat{\beta}_0 + \hat{\beta}_2 d^*\right)\right)(d - d^*)$$

$$\widehat{IIE}(d, d^*) = \hat{\beta}_2(\hat{\gamma}_3 + \hat{\gamma}_4 d)(d - d^*)$$

$$OE(d, d^*) = \widehat{IDE}(d, d^*) + \widehat{IIE}(d, d^*)$$

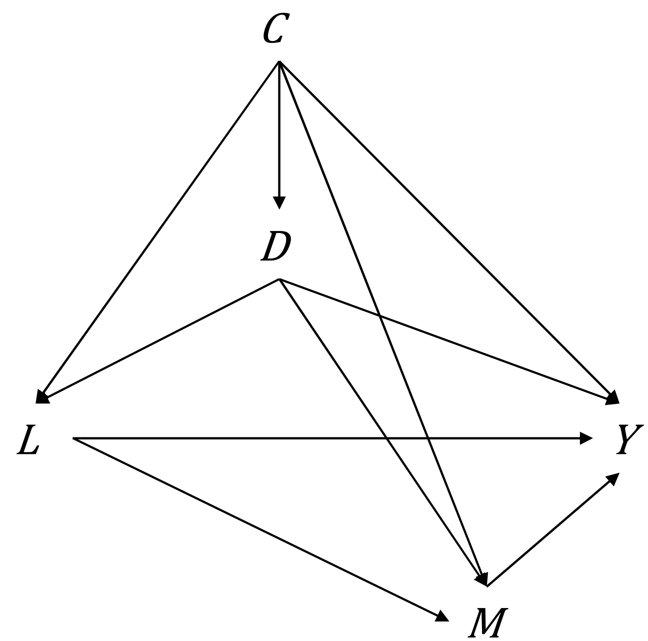
$$\widehat{CDE}(d, d^*, m) = (\hat{\gamma}_2 + \hat{\gamma}_4 m)(d - d^*)$$

- Interval estimates and p-values can be computed using the nonparametric bootstrap

RWR estimation

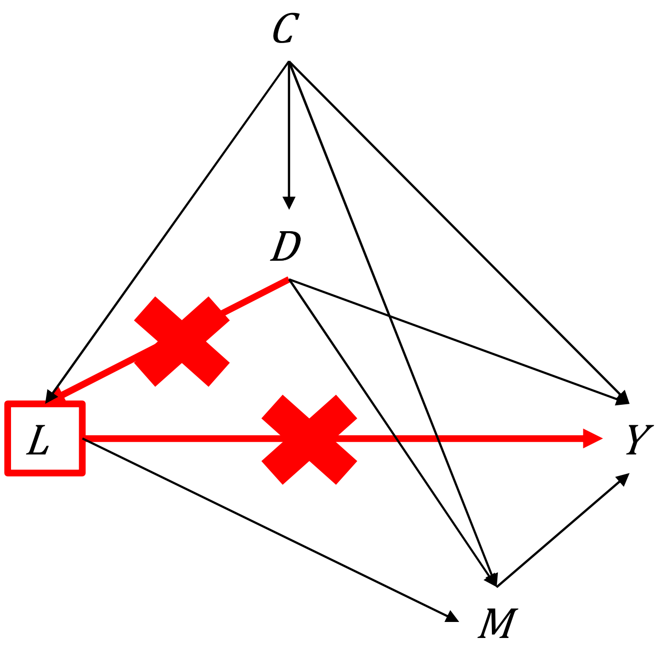
- The defining feature of RWR is that it adjusts for a residual transformation of the exposure-induced confounders

- Why adjust for these residual terms rather than the exposure-induced confounders themselves?



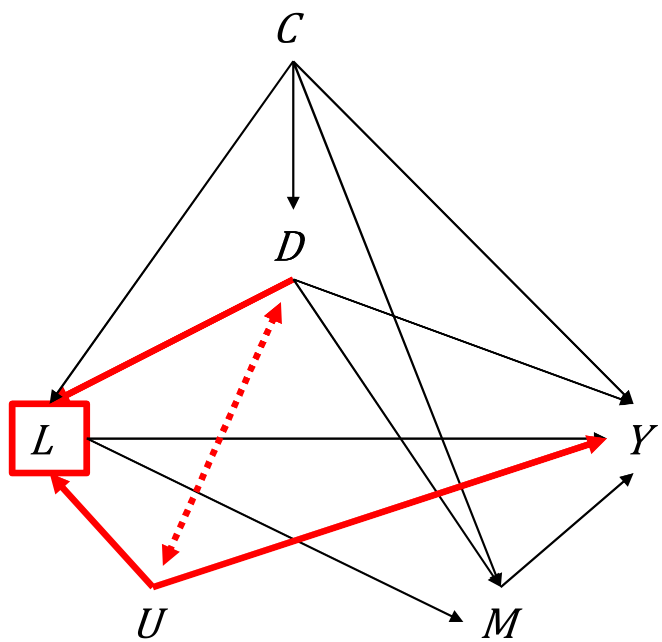
RWR estimation

- Why adjust for these residual terms rather than the exposure-induced confounders themselves?
  - Adjusting for the exposure-induced confounder L blocks the causal path $D \to L \to Y$
  - This would lead to bias in estimates of interventional effects due to "over-control of intermediate pathways"



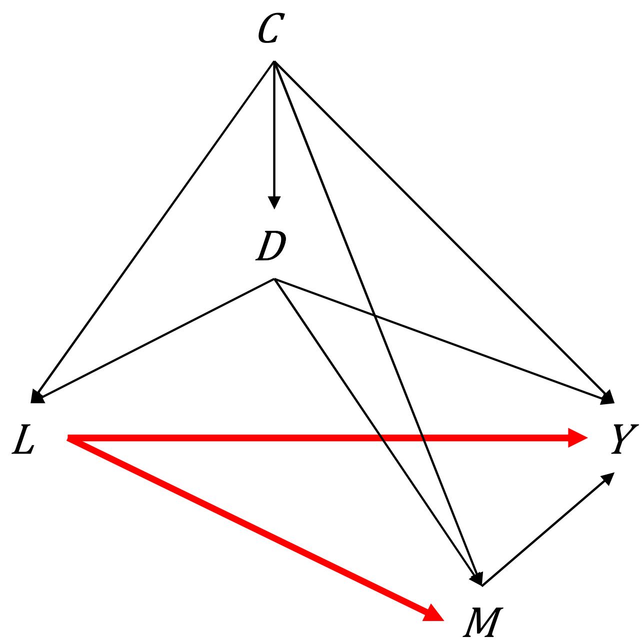
RWR estimation

- Why adjust for these residual terms rather than the exposure-induced confounders themselves?
  - Adjusting for L unblocks the non-causal path $D \rightarrow L \leftarrow U \rightarrow Y$ from the exposure to the outcome
  - This would introduce bias due to "endogenous selection" or "collider stratification"
  - L is a collider of D and U, so adjusting for L would induce a non-causal association between the exposure and unobserved variable
  - Because U also affects Y, adjusting for L would induce a non-causal association between the exposure and outcome



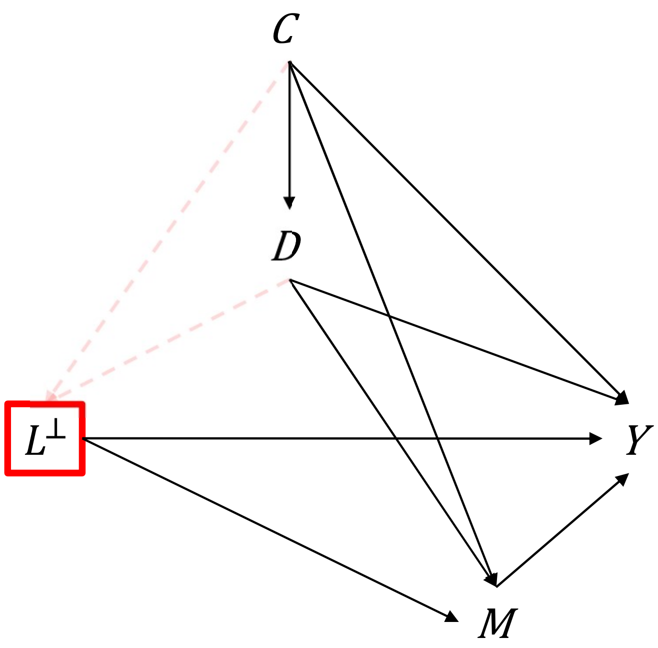
RWR estimation

- Why adjust for these residual terms rather than the exposure-induced confounders themselves?
  - At the same time, not adjusting for L leaves the non-causal path $M \leftarrow L \rightarrow Y$ from the mediator to the outcome unblocked
  - This would lead to confounding bias
  - Thus, exposure-induced confounders seemingly present a “damned if you do, damned if you don’t” dilemma



RWR estimation

- Why adjust for these residual terms rather than the exposure-induced confounders themselves?
  - RWR circumvents these problems by adjusting for a residual transformation of the exposure-induced confounders
  - Residualizing the exposure-induced confounders neutralizes the causal paths emanating from D and C into L
  - As a result, the residualized confounders can be included in an outcome regression to adjust for confounding…
  - …while sidestepping the pitfalls associated with naive confounder adjustment



Summary

- Interventional effects can be estimated using linear models for the mediator, outcome, and exposure-induced confounders fit to sample data by the method of least squares
- RWR is consistent provided that the assumptions required for identification are satisfied and the models used for estimation are correctly specified
- RWR can easily accommodate exposure-mediator interactions and effect moderation across levels of both the baseline confounders and the exposure-induced confounders
- RWR can also easily be adapted for multiple exposure-induced confounders

Example: NLSY79

- 1979 National Longitudinal Study of Youth
  - Exposure (D)
    - sample member attended college before age 22
  - Outcome (Y):
    - standardized scores on the CES-D at age 40
  - Covariates (C):
    - Race, gender, parental education, occupation, and income, household size, AFQT scores
  - A potential mediator (M)
    - household income between age 35-40
  - A potential exposure-induced confounder (L)
    - unemployment between age 35-40

Example: NLSY79

- Many studies have documented that going to college seems to reduce the likelihood of becoming depressed later in life—but how does this effect come about?
- One possibility is that a more advanced education reduces depression by increasing the income they have at their disposal
  - Does income mediate the effect of college attendance on depression?
- However, unemployment may independently affect both income and depression, and it is affected by college attendance
  - In other words, unemployment may be an exposure-induced confounder

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using RWR

```
### wk 6 nlsy tutorial ###
rm(list=ls())

## load/install libraries ##
packages<-c("dplyr", "tidyr", "foreign", "foreach", "doParallel", "doRNG", "devtools")
install.packages(packages)

for (package.i in packages) {
  suppressPackageStartupMessages(library(package.i, character.only=TRUE))
}

## load data ##
datadir <- "C:/Users/Geoff/Dropbox/D/courses/2024-25_UOFCHICAGO/SOCI_40258_CAUSAL_MEDIATION/data/"
nlsy <- read.dta(paste(datadir, "nlsy79.dta", sep=""))

Y <- "std_cesd_age40"
D <- "att22"
L <- "ever_unemp_age3539"
M <- "log_faminc_adj_age3539"
C <- c("female", "black", "hispan", "paredu", "parprof", "parinc_prank", "famsize", "afqt3")

nlsy <- nlsy[complete.cases(nlsy[,c(C,D,L,M,"cesd_age40")]),] |>
  mutate(std_cesd_age40 = (cesd_age40 - mean(cesd_age40)) / sd(cesd_age40))
```

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using RWR

```
#load R functions
source("https://raw.githubusercontent.com/causalMedAnalysis/causalMedR/refs/heads/main/utils.R")
source("https://raw.githubusercontent.com/causalMedAnalysis/causalMedR/refs/heads/main/rwrlite.R")

#install dependencies
install_github("xiangzhou09/rwrmed")

#specify form for linear models
Lformx <- ever_unemp_age3539 ~ att22 * (female + black + hispan +
  paredu + parprof + parinc_prank + famsize + afqt3)

Mformx <- log_faminc_adj_age3539 ~ att22 * (female + black + hispan +
  paredu + parprof + parinc_prank + famsize + afqt3)

Yformx <- std_cesd_age40 ~ (log_faminc_adj_age3539 * att22) *
  (female + black + hispan + paredu + parprof + parinc_prank +
  famsize + afqt3 + ever_unemp_age3539)
```

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using RWR

```r
#compute point estimates and bootstrap inferential stats
rwrest <- rwrlite(data = nlsy, D = D,  C = C, m = log(50000),
  Y_formula = Yformx, M_formula = Mformx, L_formula_list = list(Lformx),
  boot = TRUE, boot_reps = 2000, boot_seed = 60637, boot_parallel = TRUE)

rwr_boot_est <- data.frame(
  param = c("OE(1,0)", "IDE(1,0)", "IIE(1,0)", "CDE(1,0,ln(50k))"),
  est = c(rwrest$OE, rwrest$IDE, rwrest$IIE, rwrest$CDE),
  ci_lo = c(rwrest$ci_OE[1], rwrest$ci_IDE[1], rwrest$ci_IIE[1], rwrest$ci_CDE[1]),
  ci_hi = c(rwrest$ci_OE[2], rwrest$ci_IDE[2], rwrest$ci_IIE[2], rwrest$ci_CDE[2]),
  pval = c(rwrest$pvalue_OE, rwrest$pvalue_IDE, rwrest$pvalue_IIE, rwrest$pvalue_CDE)) |>
  mutate(across(.cols = !param, .fns = \(x) round(x, 3)))

print(rwr_boot_est)
```

```
> print(rwr_boot_est)
             param    est  ci_lo  ci_hi  pval
1          OE(1,0) -0.101 -0.202 -0.005 0.040
2         IDE(1,0) -0.047 -0.161  0.056 0.367
3         IIE(1,0) -0.054 -0.092 -0.009 0.016
4 CDE(1,0,ln(50k)) -0.075 -0.172  0.020 0.118
```

Limitations of RWR

- Models that are linear in the parameters may not perform very well when the mediator or outcome is binary, nominal, ordinal, or a count
  - In such applications, any linear model for the conditional expectation function is likely incorrect, potentially leading to misspecification bias
- RWR is therefore best suited to applications in which both the mediator and outcome are unbounded and possess equal-interval scaling
- Nevertheless, there are some situations where a linear model can provide a reasonable approximation for the conditional expectation of a binary, ordinal, or count variable, and thus RWR may still be defensible

Estimation via simulation

- Interventional and controlled direct effects can also be estimated using a simulation approach that is implemented with generalized linear models (GLMs)

- The class of GLMs is broad and subsumes normal linear regression as a special case, but it also includes many other nonlinear models, such as logit, probit, and Poisson regression

- This approach to estimation is therefore extremely general and can be used in a wide variety of different applications (i.e., with continuous, binary, ordinal, nominal, or count variables)

Estimation via simulation for interventional effects

- The simulation estimator for interventional effects is implemented through the following series of steps:

  1. Fit models for the mediator, outcome, and the exposure-induced confounders
  2. Simulate potential values for the exposure-induced confounders
  3. Simulate potential values for the mediator
  4. Simulate potential outcomes
  5. Compute effect estimates using the simulated outcomes

Estimation via simulation for interventional effects

- Step 1: fit models for the mediator, outcome, and the exposure-induced confounder
  - Fit a GLM for the exposure-induced confounder given the baseline confounders and exposure, denoted by $q(L|C,D)$
  - Fit a GLM for the mediator given the baseline confounders and the exposure, denoted by $g(M|C,D)$
  - Fit another GLM for the outcome given the baseline confounders, exposure, mediator, and the exposure-induced confounder, denoted by $h(Y|C,D,L,M)$
  - Let $\hat{q}(L|C,D)$, $\hat{g}(M|C,D)$, and $\hat{h}(Y|C,D,M)$ denote these models with their parameters estimated by maximum likelihood

Estimation via simulation for interventional effects

- Step 2: simulate potential values for the exposure-induced confounder
  - For every individual in the sample…
    - simulate $10^3 \leq J \leq 10^4$ copies of $L(d^*)$ from $\hat{q}(L|C, d^*)$ and then…
    - simulate another $10^3 \leq J \leq 10^4$ copies of $L(d)$ from $\hat{q}(L|C, d)$
  - Let $\tilde{L}_j(d^*)$ and $\tilde{L}_j(d)$ denote the simulated values (i.e., Monte Carlo draws) of the mediator for each simulation $j = 1, 2, \ldots, J$

Estimation via simulation for interventional effects

- Step 3: simulate potential values for the mediator
  - For every individual in the sample…
    - simulate J copies of $\mathcal{M}(d^*|C)$ from $\hat{g}(M|C, d^*)$ and then…
    - simulate another J copies of $\mathcal{M}(d|C)$ from $\hat{g}(M|C, d)$
  - Let $\widetilde{\mathcal{M}}_j(d^*|C)$ and $\widetilde{\mathcal{M}}_j(d|C)$ denote the simulated values (i.e., Monte Carlo draws) of the mediator for each simulation $j = 1, 2, \ldots, J$

Estimation via simulation for interventional effects

- Step 4: simulate potential outcomes
  - For every individual in the sample and for each simulated value of the mediator and exposure-induced confounder…
    - simulate one copy of $Y\big(d, \mathcal{M}(d|C)\big)$ from $\hat{h}\big(Y\big|C, d, \tilde{L}_j(d), \widetilde{M}_j(d|C)\big)$ and then…
    - simulate one copy of $Y\big(d^*, \mathcal{M}(d^*|C)\big)$ from $\hat{h}\big(Y\big|C, d^*, \tilde{L}_j(d^*), \widetilde{M}_j(d^*|C)\big)$ and then…
    - simulate one copy of $Y\big(d, \mathcal{M}(d^*|C)\big)$ from $\hat{h}\big(Y\big|C, d, \tilde{L}_j(d), \widetilde{M}_j(d^*|C)\big)$
  - Let $\tilde{Y}_j\big(d, \mathcal{M}(d|C)\big)$, $\tilde{Y}_j\big(d^*, \mathcal{M}(d^*|C)\big)$, and $\tilde{Y}_j\big(d, \mathcal{M}(d^*|C)\big)$ denote the simulated values (i.e., Monte Carlo draws) of the outcome for each simulation $j = 1, 2, \ldots, J$

Estimation via simulation for interventional effects

- Step 5: compute effect estimates
  - Average the difference between simulated outcomes over simulations and over sample members as follows…

$$\widehat{IDE}(d,d^*) = \frac{1}{nJ}\sum\sum_j\left[\tilde{Y}_j\left(d,\mathcal{M}(d^*|C)\right) - \tilde{Y}_j\left(d^*,\mathcal{M}(d^*|C)\right)\right]$$

$$\widehat{IIE}(d,d^*) = \frac{1}{nJ}\sum\sum_j\left[\tilde{Y}_j\left(d,\mathcal{M}(d|C)\right) - \tilde{Y}_j\left(d,\mathcal{M}(d^*|C)\right)\right]$$

$$\widehat{OE}(d,d^*) = \frac{1}{nJ}\sum\sum_j\left[\tilde{Y}_j\left(d,\mathcal{M}(d|C)\right) - \tilde{Y}_j\left(d^*,\mathcal{M}(d^*|C)\right)\right]$$

Estimation via simulation for controlled direct effects

- The simulation estimator for controlled direct effects is implemented through the following series of steps:
  1. Fit models for the exposure-induced confounders and for the outcome
  2. Simulate potential values for the exposure-induced confounders
  3. Simulate potential outcomes
  4. Compute effect estimates using the simulated outcomes

Estimation via simulation for controlled direct effects

- Step 1: fit models for the outcome and exposure-induced confounder
  - Fit a GLM for the exposure-induced confounder given the baseline confounders and exposure, denoted by $q(L|C,D)$
  - Fit another GLM for the outcome given the baseline confounders, exposure, mediator, and the exposure-induced confounder, denoted by $h(Y|C,D,L,M)$
  - Let $\hat{q}(L|C,D)$ and $\hat{h}(Y|C,D,M)$ denote these models with their parameters estimated by maximum likelihood

Estimation via simulation for controlled direct effects

- Step 2: simulate potential values for the exposure-induced confounder
  - For every individual in the sample…
    - simulate $10^3 \leq J \leq 10^4$ copies of $L(d^*)$ from $\hat{q}(L|C, d^*)$ and then…
    - simulate another $10^3 \leq J \leq 10^4$ copies of $L(d)$ from $\hat{q}(L|C, d)$
  - Let $\tilde{L}_j(d^*)$ and $\tilde{L}_j(d)$ denote the simulated values (i.e., Monte Carlo draws) of the mediator for each simulation $j = 1, 2, \ldots, J$

Estimation via simulation for controlled direct effects

- Step 3: simulate potential outcomes
  - For every individual in the sample and each simulated value of the exposure-induced confounder…
    - simulate one copy of $Y(d, m)$ from $\hat{h}\left(Y \middle| C, d, \tilde{L}_j(d), m\right)$ and then…
    - simulate one copy of $Y(d^*, m)$ from $\hat{h}\left(Y \middle| C, d^*, \tilde{L}_j(d^*), m\right)$
  - Let $\tilde{Y}_j(d, m)$ and $\tilde{Y}_j(d^*, m)$ denote the simulated values (i.e., Monte Carlo draws) of the outcome for each simulation $j = 1, 2, \ldots, J$

Estimation via simulation for controlled direct effects

- Step 4: compute effect estimates
  - Average the difference between simulated outcomes over simulations and over sample members as follows…

$$\widehat{CDE}(d, d^*, m) = \frac{1}{nJ} \sum \sum_j \left[\tilde{Y}_j(d, m) - \tilde{Y}_j(d^*, m) \right]$$

Model specification

- This approach can easily accommodate treatment-mediator interactions, covariate interactions, and nonlinear terms, as well as many different link functions and distribution models

- The steps for implementing the simulation approach are exactly the same as outlined previously, regardless of the particular form of the GLMs used for the confounders, mediator, or outcome

- The simulation estimator is therefore highly flexible

Summary

- Interventional and controlled direct effects can be estimated via simulation with a broad class of GLMs fit by the method of maximum likelihood
- These estimators are consistent provided that the assumptions required for identification are satisfied and the models used for estimation are correctly specified
- The simulation approach easily accommodates models that allow for exposure-mediator interaction, effect moderation by covariates, and nonlinearities, as well as discrete and bounded outcomes
- It can also be extended for use with multiple exposure-induced confounders

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using the simulation approach

```r
## compute estimates w/ simulation approach ##

#load R functions
source("https://raw.githubusercontent.com/causalMedAnalysis/causalMedR/refs/heads/main/medsim.R")

#specify models (using same form for linear predictors as with RWR)
specs <- list(
  list(func = "glm", formula = as.formula(Lformx), args = list(family = "binomial")),
  list(func = "lm", formula = as.formula(Mformx)),
  list(func = "lm", formula = as.formula(Yformx)))

#estimate interventional effects
ie_sim <- medsim(data = nlsy, num_sim = 1000, treatment = D, intv_med = M,
  model_spec = specs, boot = TRUE, reps = 2000, seed = 60637)

#estimate controlled direct effect
cde_sim <- medsim(data = nlsy, num_sim = 1000, treatment = D, intv_med = paste0(M,"=log(5e4)"),
  model_spec = specs, boot = TRUE, reps = 2000, seed = 60637)

```

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using the simulation approach

```r
sim_boot_est <- data.frame(
  param = c("OE(1,0)", "IDE(1,0)", "IIE(1,0)", "CDE(1,0,ln(50k))"),
  est = c(ie_sim$point.est[3], ie_sim$point.est[1], ie_sim$point.est[2], cde_sim$point.est[1]),
  ci_lo = c(ie_sim$ll.95ci[3], ie_sim$ll.95ci[1], ie_sim$ll.95ci[2], cde_sim$ll.95ci[1]),
  ci_hi = c(ie_sim$ul.95ci[3], ie_sim$ul.95ci[1], ie_sim$ul.95ci[2], cde_sim$ul.95ci[1]),
  pval = c(ie_sim$pval[3], ie_sim$pval[1], ie_sim$pval[2], cde_sim$pval[1])) |>
  mutate(across(.cols = !param, .fns = \(x) round(x, 3)))

print(sim_boot_est)
```

```
> print(sim_boot_est)
             param    est  ci_lo  ci_hi  pval
1          OE(1,0) -0.141 -0.245 -0.043 0.004
2         IDE(1,0) -0.093 -0.218  0.023 0.113
3         IIE(1,0) -0.048 -0.094 -0.001 0.045
4 CDE(1,0,ln(50k)) -0.075 -0.173  0.020 0.114
```

Limitations

- Computational complexity
  - Computing inferential statistics can be computationally demanding, leading to long processing times
- Modeling demands become more challenging with many exposure-induced confounders
  - In applications with multiple exposure-induced confounders, the method requires modeling the joint distribution of all these variables together, conditional on the baseline confounders and exposure, in order to properly simulate their potential values
  - This joint distribution can be modeled as a product of conditional distributions—one for each confounder—but correctly specifying all these models may be difficult, raising the prospect of misspecification bias

Estimation via weighting

- Weighting estimators are implemented with models for the exposure, mediator, and an exposure-induced confounder
- These models are used to construct a set of weights that transform the empirical distribution of the sample data in ways that emulate different hypothetical experiments
- The effects of interest are estimated by comparing the mean of the outcome across differently weighted samples

Estimation via weighting for interventional effects

- The weighting estimator for interventional effects is implemented through the following series of steps:
  1. Fit a model for the exposure and predict probabilities
  2. Fit a model for the exposure-induced confounder and predict probabilities
  3. Fit a model for the mediator and predict probabilities
  4. Use the predicted probabilities to construct inverse probability weights (IPWs)
  5. Compute effect estimates by comparing weighted means of the outcome

Estimation via weighting for interventional effects

- Step 1: fit a model for the exposure and predict probabilities
  - Fit a GLM for the exposure given the baseline confounders, denoted by $f(D|C)$
    - Let $\hat{f}(D|C)$ denote this model with its parameters estimated by maximum likelihood
  - For each sample member, use $\hat{f}(D|C)$ to predict…
    - the probability of exposure to d given their baseline confounders, denoted by $\hat{P}(d|C)$
    - the probability of exposure to d^* given their baseline confounders, denoted by $\hat{P}(d^*|C)$

Estimation via weighting for interventional effects

- Step 2: Fit a model for the exposure-induced confounder and predict probabilities
  - Fit a GLM for the exposure-induced confounder, denoted by $q(L|C,D)$
    - Let $\hat{q}(L|C,D)$ denote this model with its parameters estimated by maximum likelihood
  - For each sample member, use $\hat{q}(L|C,D)$ to predict…
    - the probability of each level l on the exposure-induced confounder, conditional on their observed values for the baseline confounders and exposure to d, denoted by $\hat{P}(l|C,d)$
    - the probability of each level l on the exposure-induced confounder, conditional on their observed values for the baseline confounders and exposure to d^*, denoted by $\hat{P}(l|C,d^*)$

Estimation via weighting for interventional effects

- Step 3: Fit a model for the mediator and predict probabilities
  - Fit a GLM for the mediator, denoted by $g(M|C, D, L)$
    - Let $\hat{g}(M|C, D, L)$ denote this model with its parameters estimated by maximum likelihood
  - For each sample member and every value of l, use $\hat{g}(M|C, D, L)$ to predict…
    - the probability of their observed mediator, conditional on their observed values for the baseline confounders, level l of the exposure-induced confounder, and level d of the exposure, denoted by $\hat{P}(M|C, d, l)$
    - the probability of their observed mediator, conditional on their observed values for the baseline confounders, level l of the exposure-induced confounder, and level d^* of the exposure, denoted by $\hat{P}(M|C, d^*, l)$

Estimation via weighting for interventional effects

- Step 4: construct IPWs
  - Among sample members with $D = d^*$, compute…
    - $\widehat{wt}_1 = \frac{\sum_l \hat{P}(M|C, d^*, l)\hat{P}(l|C, d^*)}{\hat{P}(d^*|C)\hat{P}(M|C, d^*, L)}$
  - Among sample members with $D = d$, compute…
    - $\widehat{wt}_2 = \frac{\sum_l \hat{P}(M|C, d, l)\hat{P}(l|C, d)}{\hat{P}(d|C)\hat{P}(M|C, d, L)}$
    - $\widehat{wt}_3 = \frac{\sum_l \hat{P}(M|C, d^*, l)\hat{P}(l|C, d^*)}{\hat{P}(d|C)\hat{P}(M|C, d, L)}$

Estimation via weighting for interventional effects

- Step 5: compute effect estimates
  - Compute differences between weighted means of the observed outcome as follows…

$$\widehat{IDE}(d, d^*) = \frac{\sum I(D=d)\widehat{wt}_3 Y}{\sum I(D=d)\widehat{wt}_3} - \frac{\sum I(D=d^*)\widehat{wt}_1 Y}{\sum I(D=d^*)\widehat{wt}_1}$$

$$\widehat{IIE}(d, d^*) = \frac{\sum I(D=d)\widehat{wt}_2 Y}{\sum I(D=d)\widehat{wt}_2} - \frac{\sum I(D=d)\widehat{wt}_3 Y}{\sum I(D=d)\widehat{wt}_3}$$

$$\widehat{OE}(d, d^*) = \frac{\sum I(D=d)\widehat{wt}_2 Y}{\sum I(D=d)\widehat{wt}_2} - \frac{\sum I(D=d^*)\widehat{wt}_1 Y}{\sum I(D=d^*)\widehat{wt}_1}$$

Estimation via weighting for interventional effects

- Step 5: compute effect estimates
  - Compute differences between weighted means of the observed outcome as follows…

$$\widehat{IDE}(d,d^*) = \frac{\sum I(D=d)\widehat{wt}_3 Y}{\sum I(D=d)\widehat{wt}_3} - \frac{\sum I(D=d^*)\widehat{wt}_1 Y}{\sum I(D=d^*)\widehat{wt}_1} = \hat{E}\left(Y\left(d,\mathcal{M}(d^*|C)\right)\right) - \hat{E}\left(Y\left(d^*,\mathcal{M}(d^*|C)\right)\right)$$

$$\widehat{IIE}(d,d^*) = \frac{\sum I(D=d)\widehat{wt}_2 Y}{\sum I(D=d)\widehat{wt}_2} - \frac{\sum I(D=d)\widehat{wt}_3 Y}{\sum I(D=d)\widehat{wt}_3} = \hat{E}\left(Y\left(d,\mathcal{M}(d|C)\right)\right) - \hat{E}\left(Y\left(d,\mathcal{M}(d^*|C)\right)\right)$$

$$\widehat{OE}(d,d^*) = \frac{\sum I(D=d)\widehat{wt}_2 Y}{\sum I(D=d)\widehat{wt}_2} - \frac{\sum I(D=d^*)\widehat{wt}_1 Y}{\sum I(D=d^*)\widehat{wt}_1} = \hat{E}\left(Y\left(d,\mathcal{M}(d|C)\right)\right) - \hat{E}\left(Y\left(d^*,\mathcal{M}(d^*|C)\right)\right)$$

Weighted pseudo-samples

- Weighting sample members with $D = d^*$ by $\widehat{wt}_1 = \frac{\sum_l \hat{P}(M|C,d^*,l)\hat{P}(l|C,d^*)}{\hat{P}(d^*|C)\hat{P}(M|C,d^*,L)}$ transforms the empirical distribution of the data so that…
  - exposure to d^* appears to have been randomly assigned
  - the mediator appears to have been randomly drawn from its distribution under exposure d^*
- Weighting sample members with $D = d$ by $\widehat{wt}_2 = \frac{\sum_l \hat{P}(M|C,d,l)\hat{P}(l|C,d)}{\hat{P}(d|C)\hat{P}(M|C,d,L)}$ transforms the empirical distribution of the data so that…
  - exposure to d appears to have been randomly assigned
  - the mediator appears to have been randomly drawn from its distribution under exposure d

Weighted pseudo-samples

- Lastly, weighting sample members with $D = d$ by $\widehat{wt}_3 = \frac{\sum_l \hat{P}(M|C,d^*,l)\hat{P}(l|C,d^*)}{\hat{P}(d|C)\hat{P}(M|C,d,L)}$ transforms the empirical distribution of the data so that…
    - exposure to d appears to have been randomly assigned
    - the mediator appears to have been randomly drawn from its distribution under the alternative exposure d^*

Estimation via weighting for controlled direct effects

- A weighting estimator for controlled direct can be implemented through the following series of steps:
  1. Fit a model for the exposure and predict probabilities
  2. Fit a model for the mediator and predict probabilities
  3. Use the predicted probabilities to construct inverse probability weights (IPWs)
  4. Compute effect estimates by comparing weighted means of the outcome

Estimation via weighting for controlled direct effects

- Step 1: fit a model for the exposure and predict probabilities
  - Fit a GLM for the exposure given the baseline confounders, denoted by $f(D|C)$
    - Let $\hat{f}(D|C)$ denote this model with its parameters estimated by maximum likelihood
  - For each sample member, use $\hat{f}(D|C)$ to predict…
    - the probability of their observed exposure given their baseline confounders, denoted by $\hat{P}(D|C)$

Estimation via weighting for controlled direct effects

- Step 2: fit a model for the mediator and predict probabilities
  - Fit a GLM for the mediator, denoted by $g(M|C, D, L)$
    - Let $\hat{g}(M|C, D, L)$ denote this model with its parameters estimated by maximum likelihood
  - For each sample member, use $\hat{g}(M|C, D, L)$ to predict…
    - the probability of their observed mediator given their observed values on both the baseline and exposure-induced confounders, denoted by $\hat{P}(M|C, D, L)$

Estimation via weighting for controlled direct effects

- Step 3: construct IPWs
  - For all sample members, compute…

$$\widehat{wt}_4 = \frac{1}{\hat{P}(M|C,D,L)\hat{P}(D|C)}$$

Estimation via weighting for controlled direct effects

- Step 4: compute effect estimates
  - Compute differences between weighted means of the observed outcome as follows…

$$\widehat{CDE}(d, d^*, m) = \frac{\sum I(D=d, M=m)\widehat{wt}_4 Y}{\sum I(D=d, M=m)\widehat{wt}_4} - \frac{\sum I(D=d^*, M=m)\widehat{wt}_4 Y}{\sum I(D=d^*, M=m)\widehat{wt}_4}$$

  - Alternatively…
    - Using weighted least squares (WLS) with weights equal to $\widehat{wt}_4$, fit a model for the outcome given the exposure and mediator only:

$\check{E}(Y|d, m) = \check{v}_0 + \check{v}_2 d + m(\check{v}_3 + \check{v}_4 d)$, where "checks" denote WLS estimates

    - Then, construct an estimate for the controlled direct effect as follows:

$$\widehat{CDE}(d, d^*, m) = (\check{v}_2 + \check{v}_4 m)(d - d^*)$$

Stabilized weights

- The inverse probability weights defined previously can yield imprecise and unstable estimates in finite samples due to their high variance

- Stabilized versions of the weights are given by the following expressions:

$$\widehat{swt}_1 = \widehat{wt}_1 \times \hat{P}(d^*)$$

$$\widehat{swt}_2 = \widehat{wt}_2 \times \hat{P}(d)$$

$$\widehat{swt}_3 = \widehat{wt}_3 \times \hat{P}(d)$$

$$\widehat{swt}_4 = \widehat{wt}_4 \times \hat{P}(M|D)P(D)$$

Censored weights

- The performance of weighting estimators can usually be improved even further by censoring the weights
- Censoring the weights involves top and bottom coding very large and very small weights, respectively, to reduce to the influence of outliers, and by extension, to improve the precision of effect estimates
- In general, the greater the degree of censoring, the more stable are the weights, and consequently, also the effect estimates based thereon, but this improved stability comes at the cost of greater systematic bias

Summary

- Interventional and controlled direct effects can be estimated via weighting with different GLMs for the exposure, mediator, and an exposure-induced confounder
- These estimators are consistent provided that the assumptions required for identification are satisfied and all the models used to construct the weights are correctly specified
- Limitations
  - Essentially impossible to use with multiple exposure-induced confounders
  - Difficult to use and often unstable with continuous or many valued exposures, mediators, or exposure-induced confounders
  - Highly sensitive to model misspecification

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using inverse probability weighting

```
#compute IPW estimates
ipwmed <- function(data, Dform, Lform, Mform) {

    df <- data

    df$id <- 1:nrow(df)

    Dmodel <- glm(Dform, data=df, family=binomial("logit"))

    Lmodel <- glm(Lform, data=df, family=binomial("logit"))

    Mmodel <- lm(Mform, data=df)

    idataD0 <- df %>% mutate(att22 = 0)
    idataD1 <- df %>% mutate(att22 = 1)

    idataD0L0 <- df %>% mutate(att22 = 0, ever_unemp_age3539 = 0)
    idataD1L0 <- df %>% mutate(att22 = 1, ever_unemp_age3539 = 0)
    idataD0L1 <- df %>% mutate(att22 = 0, ever_unemp_age3539 = 1)
    idataD1L1 <- df %>% mutate(att22 = 1, ever_unemp_age3539 = 1)

    phatD_C <- df %>%
        mutate(
            pD1_C = predict(Dmodel, newdata=df, type = "response"),
            pD0_C = 1-pD1_C,
            pD1 = mean(att22),
            pD0 = 1-pD1) %>%
        select(id, pD1_C, pD0_C, pD1, pD0)
```

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using inverse probability weighting

```
phatL_D1C <- idataD1 %>%
    mutate(
        pL1_D1C = predict(Lmodel, newdata=idataD1, type = "response"),
        pL0_D1C = 1-pL1_D1C) %>%
    select(id, pL1_D1C, pL0_D1C)

phatL_D0C <- idataD0 %>%
    mutate(
        pL1_D0C = predict(Lmodel, newdata=idataD0, type = "response"),
        pL0_D0C = 1-pL1_D0C) %>%
    select(id, pL1_D0C, pL0_D0C)

phatM_D <- df %>%
    mutate(
        pM_D1 = case_when(att22 == 1 ~
            dnorm(log_faminc_adj_age3539, mean(df$log_faminc_adj_age3539[df$att==1]),
            sd=sd(df$log_faminc_adj_age3539-(mean(df$log_faminc_adj_age3539[df$att==1])*df$att22)-
                (mean(df$log_faminc_adj_age3539[df$att==0])*(1-df$att22))))),
        pM_D0 = case_when(att22 == 0 ~
            dnorm(log_faminc_adj_age3539, mean(df$log_faminc_adj_age3539[df$att==0]),
            sd=sd(df$log_faminc_adj_age3539-(mean(df$log_faminc_adj_age3539[df$att==1])*df$att22)-
            (mean(df$log_faminc_adj_age3539[df$att==0])*(1-df$att22)))))) %>%
    select(id, pM_D1, pM_D0)
```

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using inverse probability weighting

```r
phatM_D1LC <- idataD1 %>%
    mutate(
        pM_D1LC = dnorm(log_faminc_adj_age3539,
            predict(Mmodel, newdata=idataD1, type = "response"),
            sd=sigma(Mmodel))) %>%
    select(id, pM_D1LC)

phatM_D0LC <- idataD1 %>%
    mutate(pM_D0LC = dnorm(log_faminc_adj_age3539,
        predict(Mmodel, newdata=idataD0, type = "response"),
        sd=sigma(Mmodel))) %>%
    select(id, pM_D0LC)

phatM_D0L0C <- idataD0L0 %>%
    mutate(pM_D0L0C = dnorm(log_faminc_adj_age3539,
        predict(Mmodel, newdata=idataD0L0, type = "response"),
        sd=sigma(Mmodel))) %>%
    select(id, pM_D0L0C)

phatM_D1L0C <- idataD1L0 %>%
    mutate(pM_D1L0C = dnorm(log_faminc_adj_age3539,
        predict(Mmodel, newdata=idataD1L0, type = "response"),
        sd=sigma(Mmodel))) %>%
    select(id, pM_D1L0C)

phatM_D0L1C <- idataD0L1 %>%
    mutate(pM_D0L1C = dnorm(log_faminc_adj_age3539,
        predict(Mmodel, newdata=idataD0L1, type = "response"),
        sd=sigma(Mmodel))) %>%
    select(id, pM_D0L1C)

phatM_D1L1C <- idataD1L1 %>%
    mutate(pM_D1L1C = dnorm(log_faminc_adj_age3539,
        predict(Mmodel, newdata=idataD1L1, type = "response"),
        sd=sigma(Mmodel))) %>%
    select(id, pM_D1L1C)
```

```r
df.wts <- df %>%
    full_join(phatD_C, by = "id") %>%
    full_join(phatL_D1C, by = "id") %>%
    full_join(phatL_D0C, by = "id") %>%
    full_join(phatM_D, by = "id") %>%
    full_join(phatM_D1LC, by = "id") %>%
    full_join(phatM_D0LC, by = "id") %>%
    full_join(phatM_D0L0C, by = "id") %>%
    full_join(phatM_D1L0C, by = "id") %>%
    full_join(phatM_D0L1C, by = "id") %>%
    full_join(phatM_D1L1C, by = "id")

df.wts <- df.wts %>%
    mutate(
        sw1 = case_when(att22 == 0 ~
            (pD0/pD0_C) * (1/pM_D0LC) *
            ((pM_D0L0C * pL0_D0C) + (pM_D0L1C * pL1_D0C))),
        sw2 = case_when(att22 == 1 ~
            (pD1/pD1_C) * (1/pM_D1LC) *
            ((pM_D1L0C * pL0_D1C) + (pM_D1L1C * pL1_D1C))),
        sw3 = case_when(att22 == 1 ~
            (pD1/pD1_C) * (1/pM_D1LC) *
            ((pM_D0L0C * pL0_D0C) + (pM_D0L1C * pL1_D0C))),
        sw4 = case_when(
            att22 == 1 ~ (pD1/pD1_C) * (pM_D1/pM_D1LC),
            att22 == 0 ~ (pD0/pD0_C) * (pM_D0/pM_D0LC))) %>%
    mutate(
        across(c(sw1, sw2, sw3, sw4),
        ~ifelse(. < quantile(., 0.01, na.rm = TRUE),
            quantile(., 0.01, na.rm = TRUE),
                ifelse(. > quantile(., 0.99, na.rm = TRUE),
            quantile(., 0.99, na.rm = TRUE), .))))
```

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using inverse probability weighting

```
    Ehat_Y0M0 <- weighted.mean(df.wts$std_cesd_age40[df.wts$att22==0],
        df.wts$sw1[df$att22==0])

    Ehat_Y1M1 <- weighted.mean(df.wts$std_cesd_age40[df.wts$att22==1],
        df.wts$sw2[df$att22==1])

    Ehat_Y1M0 <- weighted.mean(df.wts$std_cesd_age40[df.wts$att22==1],
        df.wts$sw3[df$att22==1])

    IDE <- Ehat_Y1M0-Ehat_Y0M0
    IIE <- Ehat_Y1M1-Ehat_Y1M0
    OE <- Ehat_Y1M1-Ehat_Y0M0

    Ymodel.wtd <- lm(std_cesd_age40 ~ att22 * log_faminc_adj_age3539,
        data=df.wts, weights=sw4)

    CDE <- Ymodel.wtd$coefficients["att22"] +
        log(50000)*Ymodel.wtd$coefficients["att22:log_faminc_adj_age3539"]

    point.est <- list(IDE, IIE, OE, CDE)

    return(point.est)
    }

#specify form of models for D, L, and M
Dform.x <- att22 ~ female + black + hispan + paredu + parprof +
    parinc_prank + famsize + afqt3

Lform.x <- ever_unemp_age3539 ~ att22 *
    (female + black + hispan + paredu + parprof + parinc_prank + famsize + afqt3)

Mform.x <- log_faminc_adj_age3539 ~ att22 *
    (ever_unemp_age3539 + female + black + hispan + paredu + parprof +
    parinc_prank + famsize + afqt3)

ipwmed.est <- ipwmed(data=nlsy, Dform=Dform.x, Lform=Lform.x, Mform=Mform.x)
ipwmed.est <- matrix(unlist(ipwmed.est), ncol=4, byrow=TRUE)
```

```
my.cluster <- parallel::makeCluster(ncores, type="PSOCK")
doParallel::registerDoParallel(cl=my.cluster)
clusterExport(cl=my.cluster,
    list("medsim", "Lform.x", "Mform.x", "Yform.x"),
    envir=environment())
clusterEvalQ(cl=my.cluster, library(dplyr))
registerDoRNG(3308004)

ipwmed.boot <- foreach(i=1:2000, .combine=cbind) %dopar% {

    boot.data <- nlsy[sample(nrow(nlsy), nrow(nlsy), replace=TRUE),]

    boot.est <- ipwmed(data=boot.data, Dform=Dform.x, Lform=Lform.x, Mform=Mform.x)

    return(boot.est)
    }


stopCluster(my.cluster)
rm(my.cluster)

ipwmed.boot <- matrix(unlist(ipwmed.boot), ncol=4, byrow=TRUE)

ipwmed.output <- matrix(data=NA, nrow=4, ncol=4)

for (i in 1:nrow(ipwmed.output)) {

    ipwmed.output[i,1] <- round(ipwmed.est[i], digits=3)
    ipwmed.output[i,2] <- round(quantile(ipwmed.boot[,i], prob=0.025), digits=3)
    ipwmed.output[i,3] <- round(quantile(ipwmed.boot[,i], prob=0.975), digits=3)
    }
```

Example: NLSY79

- Compute estimates for the interventional and controlled direct effects of college attendance on depression using inverse probability weighting

```r
IDE_null <- IIE_null <- OE_null <- CDE_null <- 0

ipwmed.boot <- as.data.frame(ipwmed.boot)

ipwmed.boot <-
    ipwmed.boot %>%
        mutate(
            IDEpval = 2*min(mean(ifelse(V1>IDE_null, 1, 0)),
                mean(ifelse(V1<IDE_null, 1, 0))),
            IIEpval = 2*min(mean(ifelse(V2>IIE_null, 1, 0)),
                mean(ifelse(V2<IIE_null, 1, 0))),
            OEpval = 2*min(mean(ifelse(V3>OE_null, 1, 0)),
                mean(ifelse(V3<OE_null, 1, 0))),
            CDEpval = 2*min(mean(ifelse(V4>CDE_null, 1, 0)),
                mean(ifelse(V4<CDE_null, 1, 0))))

for (i in 1:nrow(ipwmed.output)) {
    ipwmed.output[i,4] <- round(ipwmed.boot[1,i+4], digits=3)
    }

ipwmed.output <- data.frame(ipwmed.output, row.names=c("IDE", "IIE", "OE", "CDE"))
colnames(ipwmed.output) <- c("point.est", "ll.95ci", "ul.95ci", "pval")

print(ipwmed.output)

```

```
> print(ipwmed.output)
    point.est ll.95ci ul.95ci  pval
IDE    -0.159  -0.275  -0.021 0.023
IIE    -0.019  -0.060   0.011 0.232
OE     -0.178  -0.289  -0.051 0.005
CDE    -0.150  -0.256  -0.042 0.008
>
```