# SOCI 40258

Causal Mediation Analysis

Week 8: Path-specific Effects

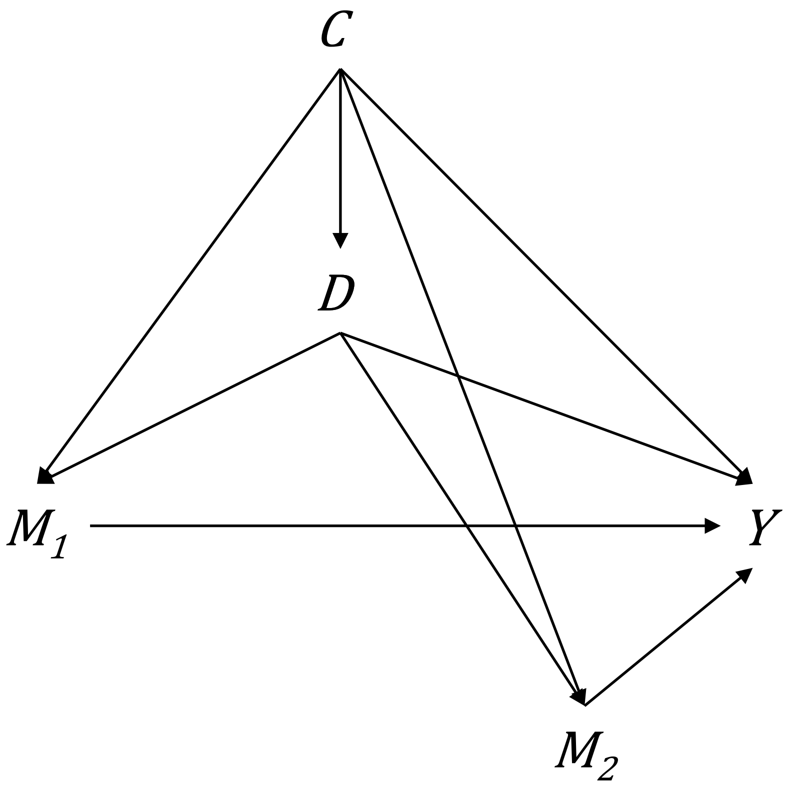
# Outline

- Graphical mediation models
- Path-specific effects
- Nonparametric identification and estimation
- Parametric estimation

# Models with multiple mediators

- In this model, the exposure $D$ affects two mediators, $M_1$ and $M_2$, which both affect the outcome $Y$

- $M_1$ does not affect $M_2$, nor does $M_2$ affect $M_1$—that is, the two mediators are causally independent

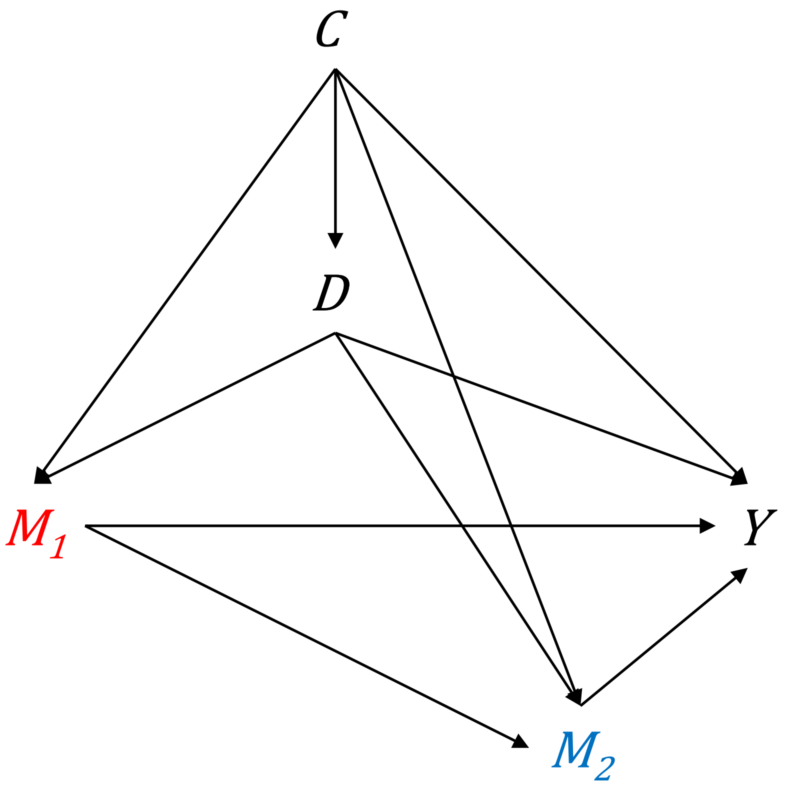


# Models with multiple mediators

- In this model, the exposure $D$ affects two mediators, $M_1$ and $M_2$, which both affect the outcome $Y$

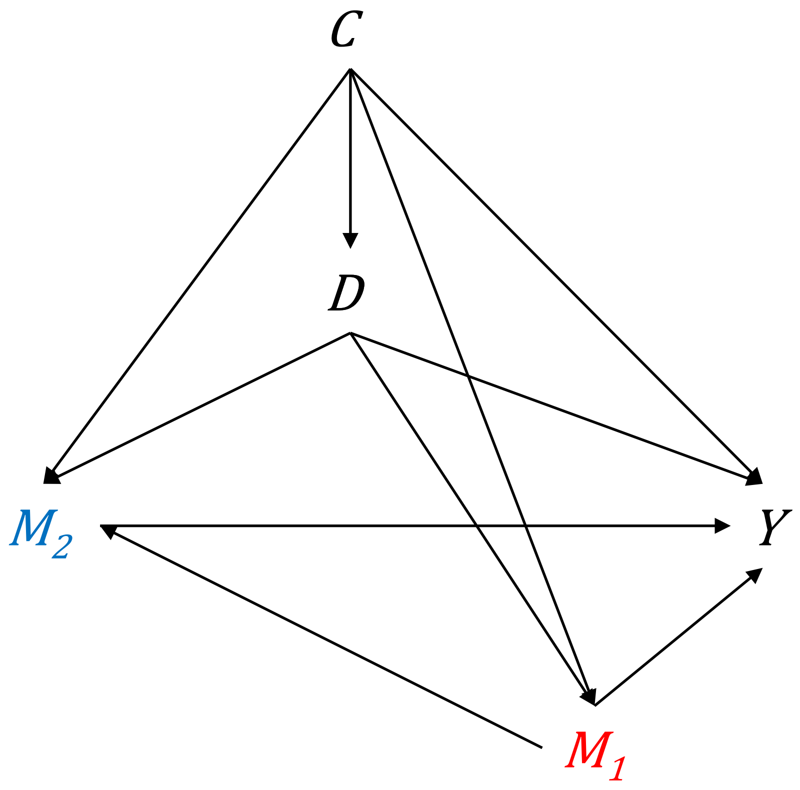
- $M_1$ now affects $M_2$, such that the mediators are causally dependent

- $M_1$ is an exposure-induced confounder with respect to the effect of $M_2$ on $Y$



# Models with multiple mediators

- In this model, the exposure $D$ affects two mediators, $M_1$ and $M_2$, which both affect the outcome $Y$
- $M_1$ affects $M_2$, such that the mediators are again causally dependent
- Let $\mathbf{M} = \{M_1, M_2, \ldots, M_K\}$ denote a set of putative mediators arranged *in causal order*



# Graphical mediation models

- The methods covered today are appropriate for data arising from a causal process resembling any of the graphical models depicted previously

- My presentation of these methods is tailored for models with two mediators and with general patterns of baseline confounding

- These methods are also appropriate for applications without any baseline confounding and with $K > 2$ mediators

# Natural effects with multiple mediators

- Natural effects with multiple mediators are very similar to the natural effects we have discussed previously, except they are defined in terms of a vector of $K$ mediators, denoted by $\mathbf{M} = \{M_1, M_2, \dots, M_K\}$

- Specifically, with multiple mediators, the average total effect of the exposure on the outcome can be decomposed into direct and indirect components as follows:

$$
\begin{aligned}
ATE(d, d^*) &= E\big(Y(d) - Y(d^*)\big) \\
&= E\left(Y\big(d, \mathbf{M}(d)\big) - Y\big(d^*, \mathbf{M}(d^*)\big)\right) \\
&= E\left(Y\big(d, \mathbf{M}(d^*)\big) - Y\big(d^*, \mathbf{M}(d^*)\big)\right) + E\left(Y\big(d, \mathbf{M}(d)\big) - Y\big(d, \mathbf{M}(d^*)\big)\right)
\end{aligned}
$$

# Natural effects with multiple mediators

- Natural effects with multiple mediators are very similar to the natural effects we have discussed previously, except they are defined in terms of a vector of $K$ mediators, denoted by $\mathbf{M} = \{M_1, M_2, \dots, M_K\}$

- Specifically, with multiple mediators, the average total effect of the exposure on the outcome can be decomposed into direct and indirect components as follows:

$$
\begin{aligned}
ATE(d, d^*) &= E\big(Y(d) - Y(d^*)\big) \\
&= E\Big(Y\big(d, \mathbf{M}(d)\big) - Y\big(d^*, \mathbf{M}(d^*)\big)\Big) \\
&= \underbrace{E\Big(Y\big(d, \mathbf{M}(d^*)\big) - Y\big(d^*, \mathbf{M}(d^*)\big)\Big)}_{\text{natural direct effect}} + \underbrace{E\Big(Y\big(d, \mathbf{M}(d)\big) - Y\big(d, \mathbf{M}(d^*)\big)\Big)}_{\text{natural indirect effect}}
\end{aligned}
$$

# Path-specific effects

- With $K = 2$ mediators arranged in causal order, the average total effect of the exposure on the outcome can also be decomposed into a set of path-specific effects as follows:

$$\begin{aligned}
ATE(d,d^*) &= E\big(Y(d) - Y(d^*)\big) \\
&= E\Big(Y\big(d,\mathbf{M}(d)\big) - Y\big(d^*,\mathbf{M}(d^*)\big)\Big) \\
&= E\Big(Y\Big(d, M_1(d), M_2\big(d, M_1(d)\big)\Big) - Y\Big(d^*, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big)\Big) \\
&= E\Big(Y\Big(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big) - Y\Big(d^*, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big)\Big) \\
&\qquad +E\Big(Y\Big(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\Big) - Y\Big(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big)\Big) \\
&\qquad +E\Big(Y\Big(d, M_1(d), M_2\big(d, M_1(d)\big)\Big) - Y\Big(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\Big)\Big)
\end{aligned}$$

# Path-specific effects

- With $K = 2$ mediators arranged in causal order, the average total effect of the exposure on the outcome can also be decomposed into a set of path-specific effects as follows:

$$
\begin{aligned}
ATE(d, d^*) &= E\big(Y(d) - Y(d^*)\big) \\
&= E\Big(Y\big(d, \mathbf{M}(d)\big) - Y\big(d^*, \mathbf{M}(d^*)\big)\Big) \\
&= E\Big(Y\Big(d, M_1(d), M_2\big(d, M_1(d)\big)\Big) - Y\Big(d^*, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big)\Big) \\
&= E\Big(Y\Big(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big) - Y\Big(d^*, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big)\Big) \quad \Big\} \; PSE_{D\to Y} \\
&\quad + E\Big(Y\Big(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\Big) - Y\Big(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\Big)\Big) \quad \Big\} \; PSE_{D\to M_2 \to Y} \\
&\quad + E\Big(Y\Big(d, M_1(d), M_2\big(d, M_1(d)\big)\Big) - Y\Big(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\Big)\Big) \quad \Big\} \; PSE_{D\to M_1 \rightsquigarrow Y}
\end{aligned}
$$

# $PSE_{D \rightarrow Y}$

- The first path-specific is formally defined as follows:

$$PSE_{D \rightarrow Y}(d, d^*) = E\left(Y\left(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right) - Y\left(d^*, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right)\right)$$

- The $PSE_{D \rightarrow Y}(d, d^*)$ is equivalent to the $MNDE(d, d^*)$
  - It represents the expected difference in the outcome if individuals had been exposed to $d$ rather than $d^*$ and if they had experienced the levels of all mediators that would have arisen naturally for them under exposure $d^*$

- It captures an effect of the exposure $D$ on the outcome $Y$ that operates through the direct causal path $D \rightarrow Y$

# $PSE_{D\rightarrow M_2\rightarrow Y}$

- The second path-specific is formally defined as follows:

$$PSE_{D\rightarrow M_2\rightarrow Y}(d,d^*) = E\left(Y\left(d,M_1(d^*),M_2\big(d,M_1(d^*)\big)\right) - Y\left(d,M_1(d^*),M_2\big(d^*,M_1(d^*)\big)\right)\right)$$

- The $PSE_{D\rightarrow M_2\rightarrow Y}(d,d^*)$ captures an effect of the exposure $D$ on the outcome $Y$ that operates through $M_2$ only

- In other words, the $PSE_{D\rightarrow M_2\rightarrow Y}(d,d^*)$ captures an effect transmitted along the $D \rightarrow M_2 \rightarrow Y$ causal path

# $PSE_{D \to M_1 \rightsquigarrow Y}$

- The second path-specific is formally defined as follows:

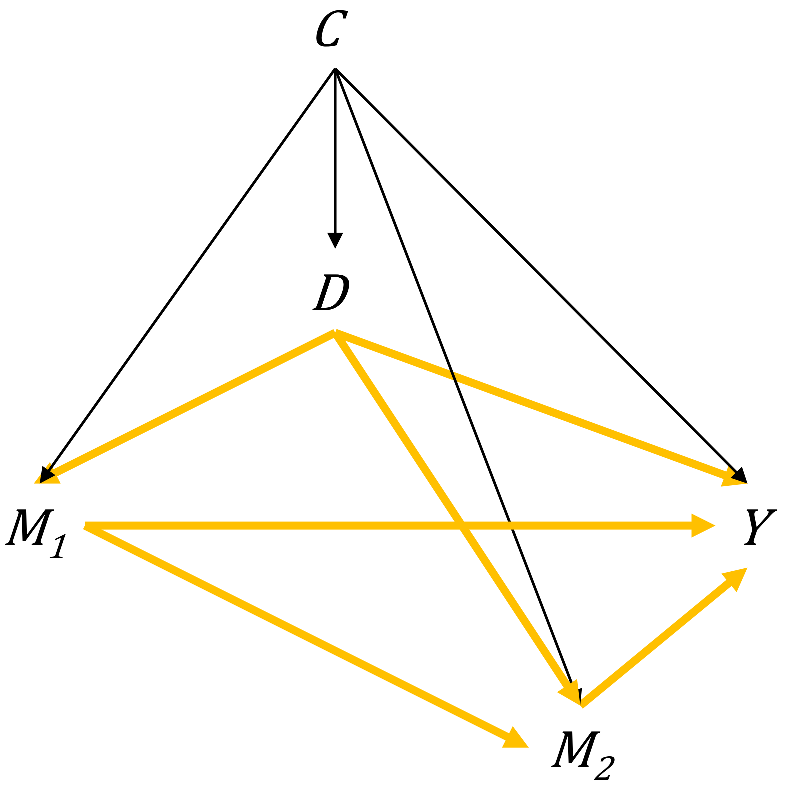
$$PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*) = E\left(Y\left(d, M_1(d), M_2\big(d, M_1(d)\big)\right) - Y\left(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\right)\right)$$

- The $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$ captures an effect of the exposure $D$ on the outcome $Y$ that operates through $M_1$
  - In other words, the $PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*)$ captures an effect transmitted along both the $D \to M_1 \to Y$ and $D \to M_1 \to M_2 \to Y$ causal paths

- The sum of the $PSE_{D \to M_2 \to Y}(d, d^*)$ and $PSE_{D \to M_2 \to Y}(d, d^*)$ is equal to the $MNIE(d, d^*)$

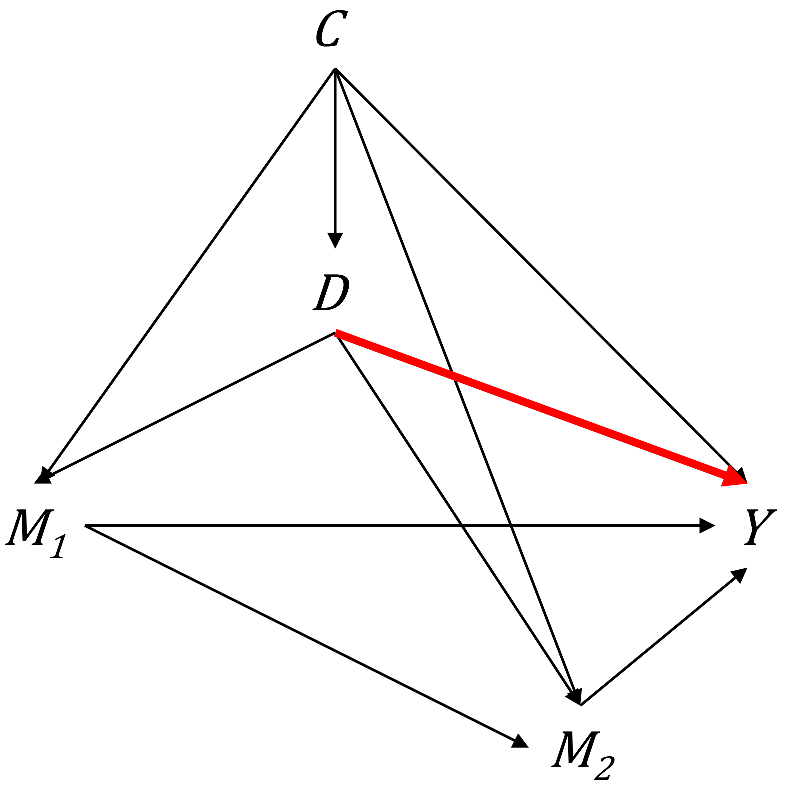
# Path-specific effects

- The total effect of $D$ on $Y$ is transmitted through the following causal paths:
  - $D \rightarrow Y$
  - $D \rightarrow M_2 \rightarrow Y$
  - $D \rightarrow M_1 \rightarrow Y$
  - $D \rightarrow M_1 \rightarrow M_2 \rightarrow Y$



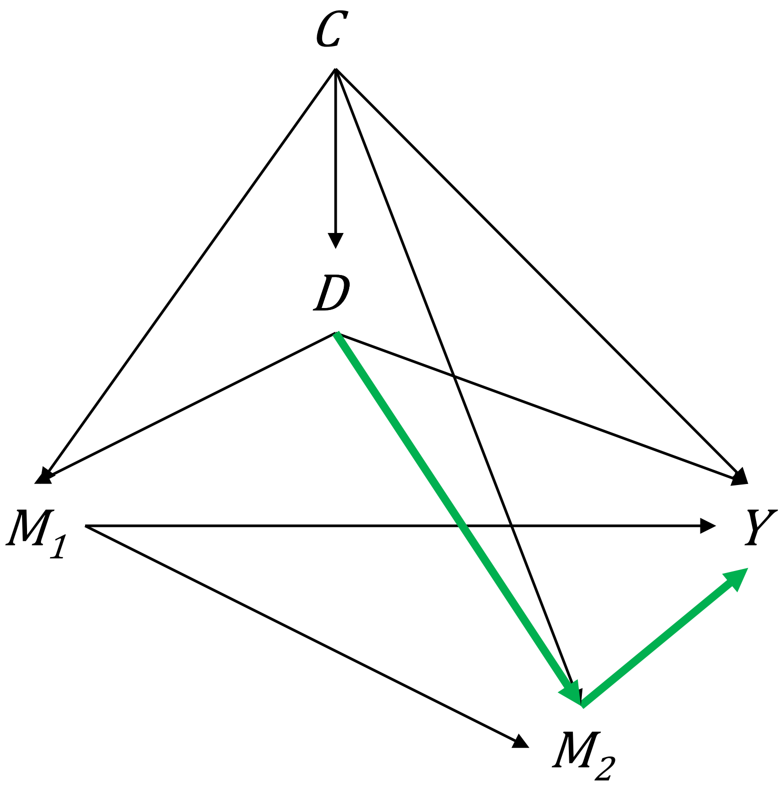
# Path-specific effects

- The total effect of $D$ on $Y$ is transmitted through the following causal paths:
  - $D \to Y$ } $PSE_{D \to Y}(d, d^*)$
  - $D \to M_2 \to Y$
  - $D \to M_1 \to Y$
  - $D \to M_1 \to M_2 \to Y$



# Path-specific effects

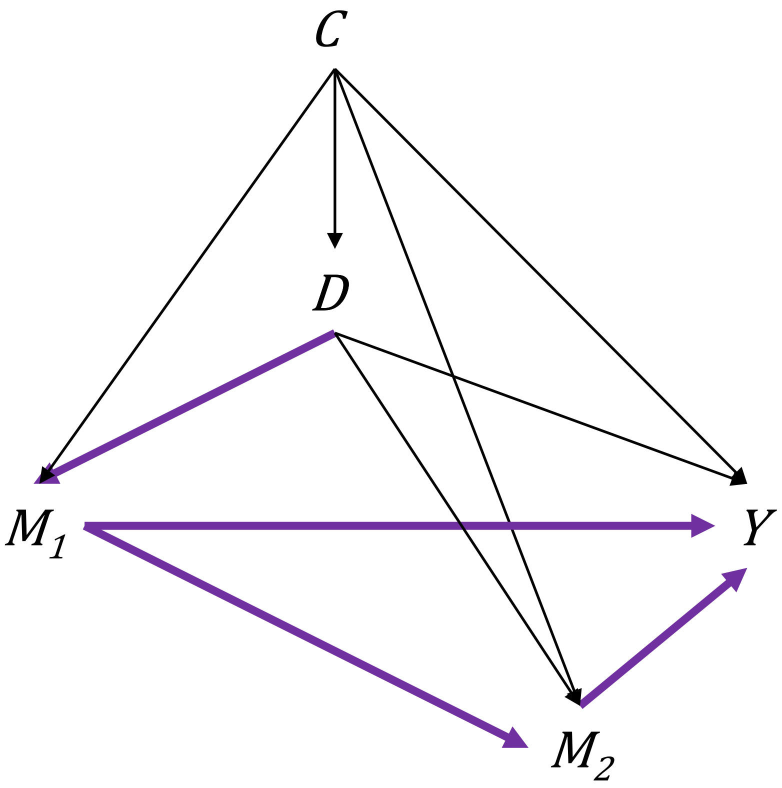
- The total effect of $D$ on $Y$ is transmitted through the following causal paths:
  - $D \to Y$
  - $D \to M_2 \to Y$ $\left.\right\}$ $PSE_{D\to M_2\to Y}(d, d^*)$
  - $D \to M_1 \to Y$
  - $D \to M_1 \to M_2 \to Y$



# Path-specific effects

- The total effect of $D$ on $Y$ is transmitted through the following causal paths:
  - $D \to Y$
  - $D \to M_2 \to Y$
  - $D \to M_1 \to Y$
  - $D \to M_1 \to M_2 \to Y$

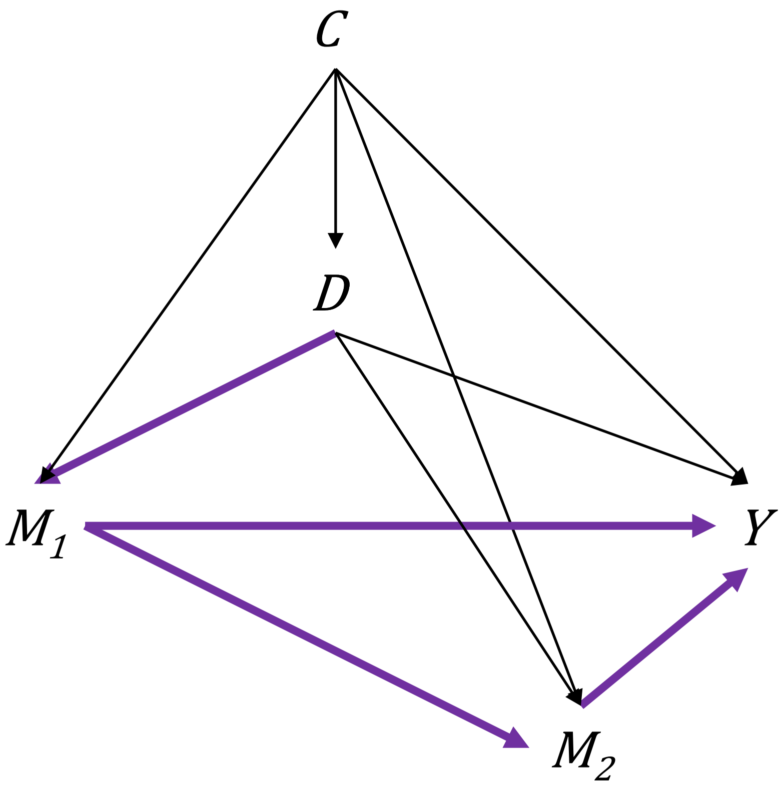
$$\left.\begin{array}{l} D \to M_1 \to Y \\ D \to M_1 \to M_2 \to Y \end{array}\right\} PSE_{D\to M_1 \rightsquigarrow Y}(d, d^*)$$



# Path-specific effects

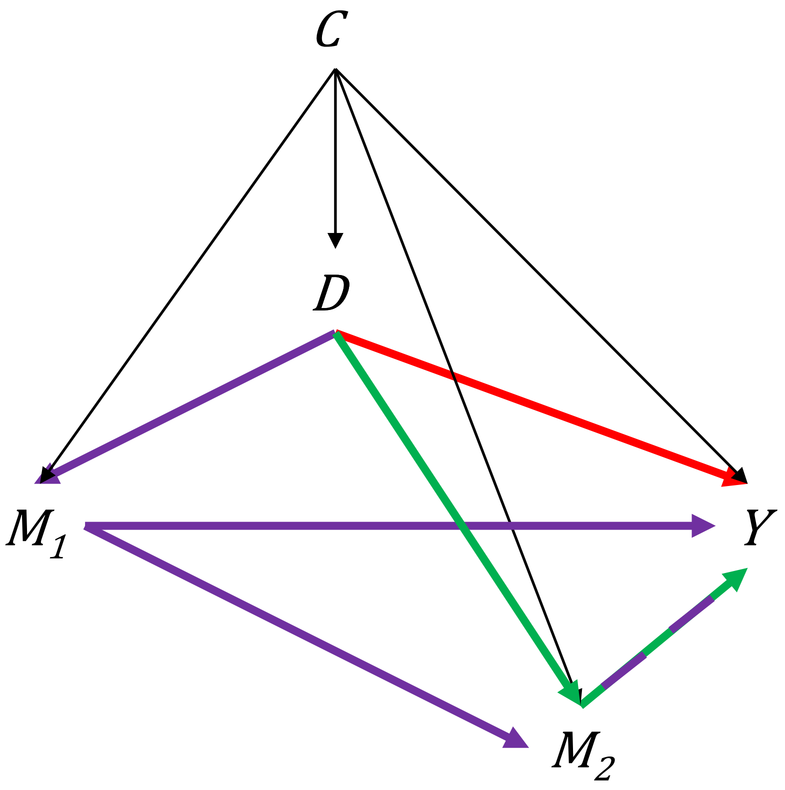
- The total effect of $D$ on $Y$ is transmitted through the following causal paths:
  - $D \rightarrow Y$
  - $D \rightarrow M_2 \rightarrow Y$
  - $D \rightarrow M_1 \rightarrow Y$
  - $D \rightarrow M_1 \rightarrow M_2 \rightarrow Y$

These paths cannot be separately identified without parametric assumptions, like linearity



# Path-specific effects

- The total effect of $D$ on $Y$ is transmitted through the following causal paths:
  - $D \rightarrow Y$
  - $D \rightarrow M_2 \rightarrow Y$
  - $D \rightarrow M_1 \rightarrow Y$
  - $D \rightarrow M_1 \rightarrow M_2 \rightarrow Y$
- In sum, path-specific effects capture the explanatory role of a focal mediator, net of all preceding mediators in a causal chain



# Path-specific effects

- By extension, the path-specific effects can also be expressed as follows:

$$PSE_{D\to Y}(d,d^*) = MNDE(d,d^*)$$

$$PSE_{D\to M_2\to Y}(d,d^*) = NDE_{M_1}(d,d^*) - MNDE(d,d^*)$$

$$PSE_{D\to M_1\rightsquigarrow Y}(d,d^*) = NIE_{M_1}(d,d^*),$$

where $NDE_{M_1}(d,d^*)$ and $NIE_{M_1}(d,d^*)$ represent the univariate natural direct and indirect effects through the first mediator $M_1$ and $MNDE(d,d^*)$ is defined as before

# Path-specific effects

- In general, path-specific effects through $k = 1, \ldots, K$ causally ordered mediators can be expressed as follows:

$$PSE_{D\to Y}(d, d^*) = NDE_{\mathbf{M}_K}(d, d^*)$$

$$PSE_{D\to M_k \rightsquigarrow Y}(d, d^*) = NDE_{\mathbf{M}_{k-1}}(d, d^*) - NDE_{\mathbf{M}_k}(d, d^*)$$

$$PSE_{D\to M_1 \rightsquigarrow Y}(d, d^*) = NIE_{M_1}(d, d^*),$$

where $NIE_{M_1}(d, d^*)$ represents the univariate indirect effect through the first mediator $M_1$ and $NDE_{\mathbf{M}_k}(d, d^*)$ represents the multivariate natural direct effect that does not operate through $\mathbf{M}_k = \{M_1, \ldots, M_k\}$

# Nonparametric identification

- Path-specific effects with multiple mediators can be nonparametrically identified if the following conditions are met for all values of $d, d', d'', m_1, m_1',$ and $m_2$:

Assumption PSE.1: $\{M_1(d'), M_2(d'', m_1), Y(d, m_1, m_2)\} \perp D|C$

Assumption PSE.2: $\{M_2(d'', m_1), Y(d, m_1, m_2)\} \perp M_1(d')|C, D$

Assumption PSE.3: $Y(d, m_1, m_2) \perp M_2(d'', m_1')|C, D, M_1$

Assumption PSE.4: $P(d|c) > 0$, $P(d|c, m_1) > 0$ and $P(d|c, m_1, m_2) > 0$

Assumption PSE.5: $M_1 = M_1(D)$, $M_2 = M_2(D, M_1)$, $Y = Y(D, M_1, M_2)$

# No unobserved exposure-outcome or exposure-mediator confounding

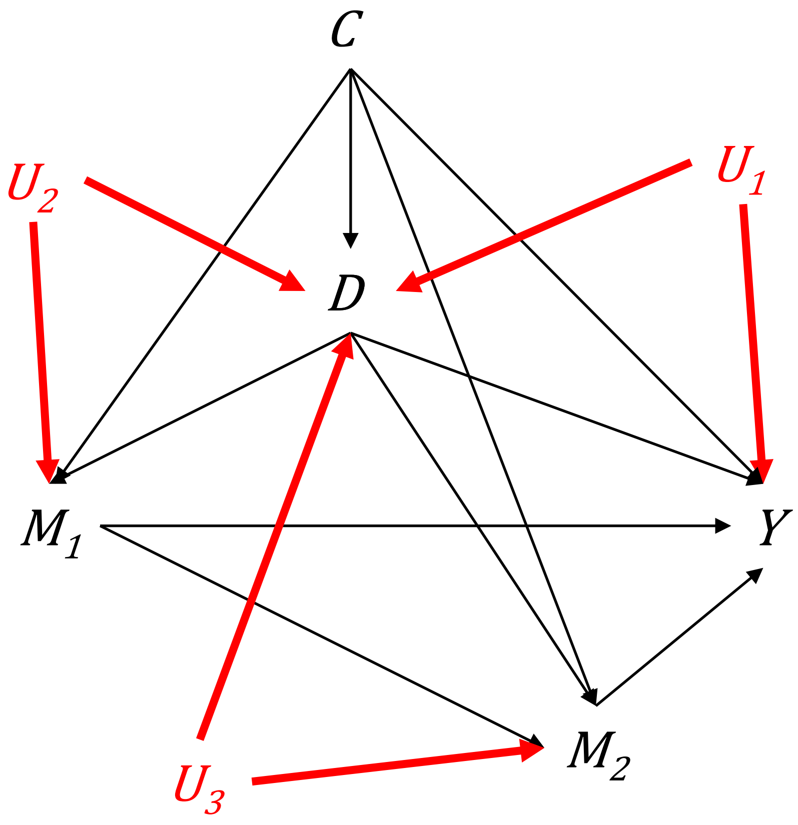
- Assumption PSE.1:

$$\{M_1(d'), M_2(d'', m_1), Y(d, m_1, m_2)\} \perp D|C$$

- This assumption requires that the exposure $D$ must be statistically independent of the potential outcomes and the potential values of all mediators, conditional on the baseline confounders $C$
- Substantively, this assumption requires that there must not be any unobserved factors that confound the exposure-outcome relationship or the relationships of the exposure with any of the mediators

# No unobserved exposure-outcome or exposure-mediator confounding

- Assumption PSE.1 would be violated if an unobserved variable jointly affects…
  - the exposure and outcome or…
  - the exposure and either mediator



# No unobserved mediator-mediator or mediator-outcome confounding

- Assumption PSE.2:

$$\{M_2(d'', m_1), Y(d, m_1, m_2)\} \perp M_1(d') | C, D$$

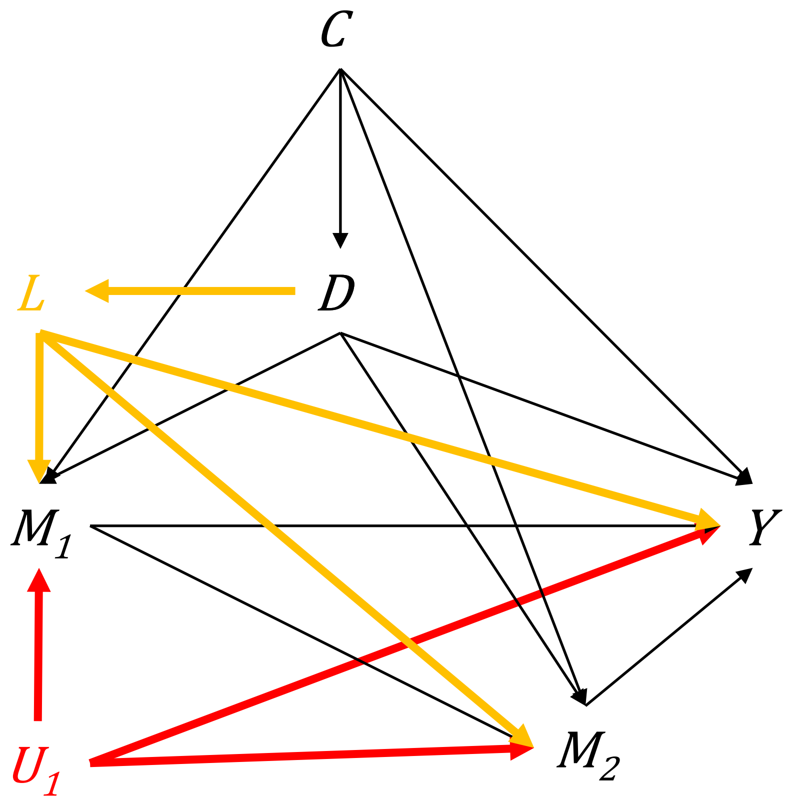
- This assumption requires that the potential values of the first mediator must be independent of the potential values for the second mediator and the outcome, conditional on the baseline confounders and exposure

- Substantively, this assumption requires that there must not be any unobserved or exposure-induced confounding for the relationship of the first mediator with the second mediator or the outcome

# No unobserved mediator-mediator or mediator-outcome confounding

- Assumption PSE.2 would be violated if an unobserved variable jointly affects $M_1$ and $M_2$ or $Y$

- It would also be violated if there were any exposure-induced confounders for the $M_1 \rightarrow M_2$ or $M_1 \rightarrow Y$ relationships…
  - …whether they are observed or not



# No unobserved or exposure-induced mediator-outcome confounding

- Assumption PSE.3:

$$Y(d, m_1, m_2) \perp M_2(d'', m_1') | C, D, M_1$$

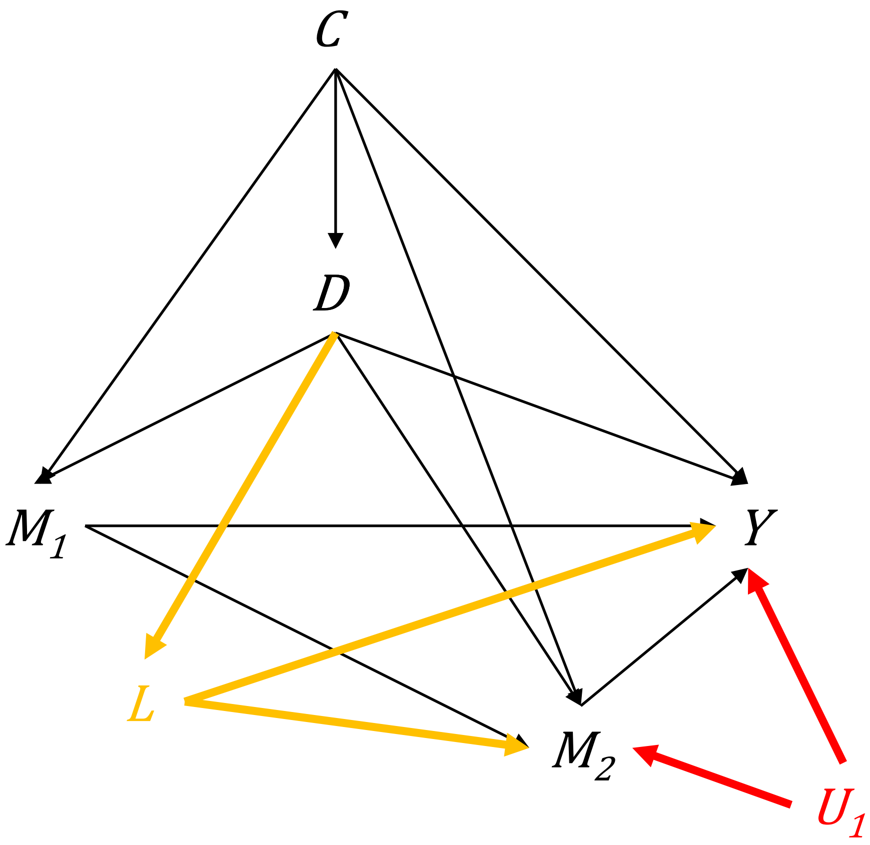
- This assumption requires that the potential values of the second mediator must be independent of the potential outcomes, conditional on the baseline confounders, exposure, and the first mediator

- Substantively, this assumption requires that there must not be any unobserved or exposure-induced confounding for the relationship of the second mediator with the outcome

# No unobserved or exposure-induced mediator-outcome confounding

- Assumption PSE.3 would be violated if an unobserved variable jointly affects $M_2$ and $Y$

- It would also be violated if there were any exposure-induced confounders for the $M_2 \rightarrow Y$ relationship…
  - …whether observed or not



# Identification formula for $PSE_{D\to Y}$

- Under assumptions PSE.1 to PSE.5, the path-specific effect from exposure directly to the outcome can be equated with a function of observable data

- Nonparametric identification formula for the $PSE_{D\to Y}$:

$$PSE_{D\to Y}(d,d^*) = E\left(Y\left(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right) - Y\left(d^*, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right)\right)$$

$$= \textstyle\sum_c \sum_{m_1} \sum_{m_2} [E(Y|c,d,m_1,m_2) - E(Y|c,d^*,m_1,m_2)]$$

$$\times P(m_2|c,d^*,m_1)P(m_1|c,d^*)P(c)$$

# Identification formula for $PSE_{D\to M_2\to Y}$

- Under assumptions PSE.1 to PSE.5, the path-specific effect from exposure to the outcome through the second mediator only can also be equated with a function of observable data

- Nonparametric identification formula for the $PSE_{D\to M_2\to Y}$:

$$PSE_{D\to M_2\to Y}(d,d^*) = E\left(Y\left(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\right) - Y\left(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right)\right)$$

$$= \sum_c \sum_{m_1} \sum_{m_2} E(Y|c,d,m_1,m_2)$$

$$\times [P(m_2|c,d,m_1) - P(m_2|c,d^*,m_1)]P(m_1|c,d^*)P(c)$$

# Identification formula for $PSE_{D\rightarrow M_1 \rightsquigarrow Y}$

- Under assumptions PSE.1 to PSE.5, the path-specific effect from exposure to the outcome through the first mediator can be equated with a function of observable data as well

- Nonparametric identification formula for the $PSE_{D\rightarrow M_1 \rightsquigarrow Y}$:

$$PSE_{D\rightarrow M_1 \rightsquigarrow Y}(d,d^*) = E\left(Y\left(d, M_1(d), M_2\big(d, M_1(d)\big)\right) - Y\left(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\right)\right)$$

$$= \sum_c \sum_{m_1} \sum_{m_2} E(Y|c,d,m_1,m_2)$$

$$\times P(m_2|c,d,m_1)[P(m_1|c,d) - P(m_1|c,d^*)]P(c)$$

# Nonparametric estimation

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

- With multiple mediators, these challenges associated with nonparametric estimation—sparsity, the curse of dimensionality, and high variance—usually preclude nonparametric estimation

- Thus, we will focus exclusively on parametric approaches to estimation, all of which are based on the identification formulas outlined previously

# Parametric estimation with linear models: a special case

- Consider the following set of linear and additive models, where $c^{\perp} = c - \bar{C}$:

$$E(M_1|c, d) = \beta_{01} + \beta_{11}^T c^{\perp} + \beta_{21} d$$

$$E(M_2|c, d, m_1) = \beta_{02} + \beta_{12}^T c^{\perp} + \beta_{22} d + \beta_{32} m_1$$

$$E(Y|c, d, m_1, m_2) = \gamma_0 + \gamma_1^T c^{\perp} + \gamma_2 d + \gamma_{31} m_1 + \gamma_{32} m_2$$

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

$$PSE_{D \to Y}(d, d^*) = \gamma_2 (d - d^*)$$

$$PSE_{D \to M_2 \to Y}(d, d^*) = \beta_{22} \gamma_{32} (d - d^*)$$

$$PSE_{D \to M_1 \rightsquigarrow Y}(d, d^*) = \beta_{21} (\gamma_{31} + \beta_{32} \gamma_{32})(d - d^*)$$

# Parametric estimation with linear models: a special case

- Path-specific effects can be estimated using linear and additive models for the mediators and outcome fit to sample data by the method of least squares

- These estimators are consistent provided that the assumptions required for identification are satisfied and provided that all the models used for estimation are correctly specified

- Essentially identical to classical linear path analysis

- Limitation: doesn’t accommodate interactions

# Parametric estimation with linear models and weighting

- We have already learned how to use linear models and weighting to estimate both univariate and multivariate natural effects

- Recall that path-specific effects are functions of particular combinations of natural effects

- Therefore, we can apply the methods discussed earlier to estimate the relevant natural effects and then use them to construct the desired path-specific effects

- This approach allows us to accommodate many different types of interactions and nonlinearities

# Estimation via simulation

- Path-specific 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 a number of nonlinear models, such as logit, probit, and Poisson regression, among others

- This approach to estimation is therefore extremely general and can be used in a wide variety of different applications

# Estimation via simulation

- The simulation estimator is implemented through a series of steps:
  1. Fit a model for the first mediator and simulate potential values
  2. Fit a model for the second mediator and simulate potential values
  3. Fit a model for the outcome
  4. Simulate potential outcomes
  5. Compute effect estimates using the simulated outcomes

# Estimation via simulation

- Step 1: fit a model for the first mediator and simulate potential values
  - Fit a GLM for $M_1$ given the baseline confounders and exposure, denoted by $g_1(M_1|C,D)$
    - Let $\hat{g}_1(M_1|C,D)$ denote these models with their parameters estimated by maximum likelihood
  - Then, for every individual in the sample…
    - Simulate one copy of $M_1(d^*)$ from $\hat{g}_1(M_1|C,d^*)$, and then simulate one copy of $M_1(d)$ from $\hat{g}_1(M_1|C,d)$
    - Repeat this step $10^3 \leq J \leq 10^4$ times
  - Let $\widetilde{M}_{j1}(d^*)$ and $\widetilde{M}_{j1}(d)$ denote the simulated values of the first mediator for each simulation $j = 1, 2, \ldots, J$

# Estimation via simulation

- Step 2: fit a model for the second mediator and simulate potential values
  - Fit a GLM for $M_2$ given the baseline confounders, exposure, and first mediator, denoted by $g_2(M_2|C, D, M_1)$; let $\hat{g}_2(M_2|C, D, M_1)$ denote this model fit by maximum likelihood
  - For every sample member and each simulated value of the first mediator…
    - Simulate one copy of $M_2\big(d^*, M_1(d^*)\big)$ from $\hat{g}_2\big(M_2\big|C, d^*, \widetilde{M}_{j1}(d^*)\big)$
    - Simulate one copy of $M_2\big(d, M_2(d)\big)$ from $\hat{g}_2\big(M_2\big|C, d, \widetilde{M}_{j1}(d)\big)$
    - Simulate one copy of $M_2\big(d, M_1(d^*)\big)$ from $\hat{g}_2\big(M_2\big|C, d, \widetilde{M}_{j1}(d^*)\big)$
  - Let $\widetilde{M}_{j2}\big(d^*, M_1(d^*)\big)$, $\widetilde{M}_{j2}\big(d, M_1(d)\big)$, and $\widetilde{M}_{j2}\big(d, M_1(d^*)\big)$ denote the simulated values for each simulation $j\ =\ 1, 2, \ldots, J$

# Estimation via simulation

- Step 3: fit a model for the outcome
  - Fit a GLM for the outcome given the baseline confounders, the exposure, and both mediators, denoted by $h(Y|C,D,M_1,M_2)$
  - Let $\hat{h}(Y|C,D,M_1,M_2)$ denote this model with its parameters estimated by maximum likelihood

# Estimation via simulation

- Step 4: simulate potential outcomes
  - For every sample member and each set of simulated mediators…
    - simulate one copy of $Y\left(d, M_1(d), M_2\big(d, M_1(d)\big)\right)$ from $\hat{h}\big(Y \big| C, d, \widetilde{M}_{j1}(d), \widetilde{M}_{j2}\big(d, M_1(d)\big)\big)$
    - simulate one copy of $Y\left(d^*, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right)$ from $\hat{h}\big(Y \big| C, d^*, \widetilde{M}_{j1}(d^*), \widetilde{M}_{j2}\big(d^*, M_1(d^*)\big)\big)$
    - simulate one copy of $Y\left(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right)$ from $\hat{h}\big(Y \big| C, d, \widetilde{M}_{j1}(d^*), \widetilde{M}_{j2}\big(d^*, M_1(d^*)\big)\big)$
    - simulate one copy of $Y\left(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\right)$ from $\hat{h}\big(Y \big| C, d, \widetilde{M}_{j1}(d^*), \widetilde{M}_{j2}\big(d, M_1(d^*)\big)\big)$
  - Let $\tilde{Y}_j\left(d, M_1(d'), M_2\big(d'', M_1(d''')\big)\right)$ denote the simulated values of the outcome for each simulation $j = 1, 2, \ldots, J$ and each value of $d$, $d'$, $d''$, and $d'''$

# Estimation via simulation

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

$$\widehat{PSE}_{D\to Y}(d,d^*)=\frac{1}{nJ}\sum\sum_j\left[\tilde{Y}_j\left(d,M_1(d^*),M_2\big(d^*,M_1(d^*)\big)\right)-\tilde{Y}_j\left(d^*,M_1(d^*),M_2\big(d^*,M_1(d^*)\big)\right)\right]$$

$$\widehat{PSE}_{D\to M_2\to Y}(d,d^*)=\frac{1}{nJ}\sum\sum_j\left[\tilde{Y}_j\left(d,M_1(d^*),M_2\big(d,M_1(d^*)\big)\right)-\tilde{Y}_j\left(d,M_1(d^*),M_2\big(d^*,M_1(d^*)\big)\right)\right]$$

$$\widehat{PSE}_{D\to M_1\rightsquigarrow Y}(d,d^*)=\frac{1}{nJ}\sum\sum_j\left[\tilde{Y}_j\left(d,M_1(d),M_2\big(d,M_1(d)\big)\right)-\tilde{Y}_j\left(d,M_1(d^*),M_2\big(d,M_1(d^*)\big)\right)\right]$$

# Model specification

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

- The steps outlined previously proceed exactly the same, regardless of the particular form of the GLMs used for the mediators and outcome

- What matters is that these models are correctly specified, or more realistically, that they are not badly misspecified

# Summary

- Path-specific effects can be estimated via simulation with a broad class of GLMs fit to sample data by the method of maximum likelihood

- These estimators are consistent provided that the assumptions required for identification are satisfied and provided that the models used for estimation are correctly specified

- Limitations
  - The method requires correctly specified models for all the mediators in causal order and the outcome, which may be difficult to achieve in practice, especially in applications with many mediators

# Regression imputation

- Path-specific effects can also be estimated using a regression imputation approach

- Unlike the other approaches we've considered, regression imputation does not require models for the mediators or for the exposure; rather, it only requires a series of models for the outcome

- Because it only requires models for the outcome, this approach to estimation is well-suited for applications with multiple mediators

# Regression imputation

- Regression imputation is implemented through a series of steps:

1. Fit a model for the outcome given the exposure and baseline confounders
    - Impute outcomes from this model under $D = d$ and $D = d^*$

2. Fit another model for the outcome given the exposure, confounders, and both mediators
    - Predict outcomes from this model under $D = d$
    - Fit a model for these predicted outcomes, and then impute outcomes under $D = d^*$

3. Fit another model for the outcome given the exposure, confounders, and first mediator only
    - Predict outcomes from this model under $D = d$
    - Fit a model for these predicted outcomes, and then impute outcomes under $D = d^*$

4. Compute effect estimates using the different imputed outcomes

# Regression imputation

- Step 1: Fit a model for the outcome given the exposure and baseline confounders, and construct imputations
  - Fit a model for the outcome given the baseline confounders and the exposure, denoted by $q(Y|C,D)$
    - Let $\hat{q}(Y|C,D)$ denote this model with its parameters estimated by least squares or maximum likelihood
  - Impute potential outcomes under $d^*$ by setting $D=d^*$ for all sample members and computing predicted values from the fitted model, denoted by $\hat{Y}(d^*)$
  - Impute potential outcomes under $d$ by setting $D=d$ for all sample members and computing predicted values from the fitted model, denoted by $\hat{Y}(d)$

# Regression imputation

- Step 2.1: fit another model for the outcome given the exposure, confounders, and both mediators
  - Fit a model for the outcome given the baseline confounders, the exposure, and both mediators, denoted by $h_1\ (Y|C, D, M_1, M_2)$
    - Let $\hat{h}_1(Y|C, D, M_1, M_2)$ denote this model with its parameters estimated by least squares or maximum likelihood
    - Predict outcomes under $d$, $M_1(D)$, and $M_2\big(D, M_1(D)\big)$ by setting $D = d$ for all sample members, leaving the mediators at their observed values, and then computing fitted values, denoted as $\hat{Y}\Big(d, M_1(D), M_2\big(D, M_1(D)\big)\Big)$

# Regression imputation

- Step 2.2: fit a model for the predicted outcomes given the exposure and confounders only
  - Next, fit a model for the predicted outcomes given only the baseline confounders and the exposure, denoted by $\tau\left(\hat{Y}\left(d, M_1(D), M_2\big(D, M_1(D)\big)\right) \middle| C, D\right)$
    - Let $\hat{\tau}\left(\hat{Y}\left(d, M_1(D), M_2\big(D, M_1(D)\big)\right) \middle| C, D\right)$ denote this model with its parameters estimated by least squares or maximum likelihood
    - Predict outcomes under $d$, $M_1(d^*)$, and $M_2\big(d^*, M_1(d^*)\big)$ by setting $D = d^*$ for all sample members and computing fitted values, denoted as $\hat{Y}\left(d, M_1(d^*), M_2\big(d^*, M_1(d^*)\big)\right)$

# Regression imputation

- Step 3.1: fit another model for the outcome given the exposure, confounders, and first mediator only
  - Fit a model for the outcome given the baseline confounders, the exposure, and the first mediator only, denoted by $h_2(Y|C,D,M_1)$
    - Let $\hat{h}_2(Y|C,D,M_1\ )$ denote this model with its parameters estimated by least squares or maximum likelihood
    - Predict outcomes under $d$, $M_1(D)$, and $M_2\big(d,M_1(D)\big)$ by setting $D=d$ for all sample members, leaving the first mediator at its observed value, and computing fitted values, denoted as $\hat{Y}\left(d,M_1(D),M_2\big(d,M_1(D)\big)\right)$

# Regression imputation

- Step 3.2: fit a model for the predicted outcomes given the exposure and confounders only
  - Next, fit a model for the predicted outcomes given the baseline confounders and exposure, denoted by $\phi\left(\hat{Y}\left(d, M_1(D), M_2\big(d, M_1(D)\big)\right)\middle|C, D\right)$
    - Let $\hat{\phi}\left(\hat{Y}\left(d, M_1(D), M_2\big(d, M_1(D)\big)\right)\middle|C, D\right)$ denote this model with its parameters estimated by least squares or maximum likelihood
    - Predict outcomes under $d$, $M_1(d^*)$, and $M_2\big(d, M_1(d^*)\big)$ by setting $D = d^*$ for all sample members and computing fitted values, denoted as $\hat{Y}\left(d, M_1(d^*), M_2\big(d, M_1(d^*)\big)\right)$

# Regression imputation

- Step 4: compute effect estimates
  - Compute differences between means of the imputed outcomes as follows…

$$\widehat{PSE}_{D\to Y}(d,d^*) = \frac{1}{n}\sum\left[\hat{Y}\left(d, M_1(d^*), M_2\left(d^*, M_1(d^*)\right)\right) - \hat{Y}(d^*)\right]$$

$$\widehat{PSE}_{D\to M_2\to Y}(d,d^*) = \frac{1}{n}\sum\left[\hat{Y}\left(d, M_1(d^*), M_2\left(d, M_1(d^*)\right)\right) - \hat{Y}\left(d, M_1(d^*), M_2\left(d^*, M_1(d^*)\right)\right)\right]$$

$$\widehat{PSE}_{D\to M_1\rightsquigarrow Y}(d,d^*) = \frac{1}{n}\sum\left[\hat{Y}(d) - \hat{Y}\left(d, M_1(d^*), M_2\left(d, M_1(d^*)\right)\right)\right]$$

# Summary

- Path-specific effects can be estimated via regression imputation with a series of different models for the outcome

- These estimators are consistent provided that the assumptions required for identification are satisfied and provided that the models used for estimation are correctly specified

- Regression imputation is especially useful in analyses of path-specific effects because this approach obviates the need to fit any models for multiple different mediators

# Generalization to $K$ mediators

- Although we've focused on implementation with only 2 mediators, all the methods covered previously are easily generalized to applications with $K > 2$ mediators

- In applications with $K > 2$ mediators…
  - estimation with linear models or weighting is adapted by estimating additional multivariate natural effects, which are then used to compute the path-specific effects of interest;
  - the simulation approach is adapted by fitting additional models for each mediator, adding terms as appropriate to the models for the mediators and outcome, and modifying the ancestral sampling procedure accordingly; and
  - the regression imputation approach is modified by adding additional steps where each mediator is successively dropped from the outcome model to produce predictions, which are then used to impute the relevant cross-world potential outcomes

# 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
  - Potential mediators ($\mathbf{M}$)
    - unemployment between age 35-40 ($M_1$)
    - household income between age 35-40 ($M_2$)

# Example: NLSY79

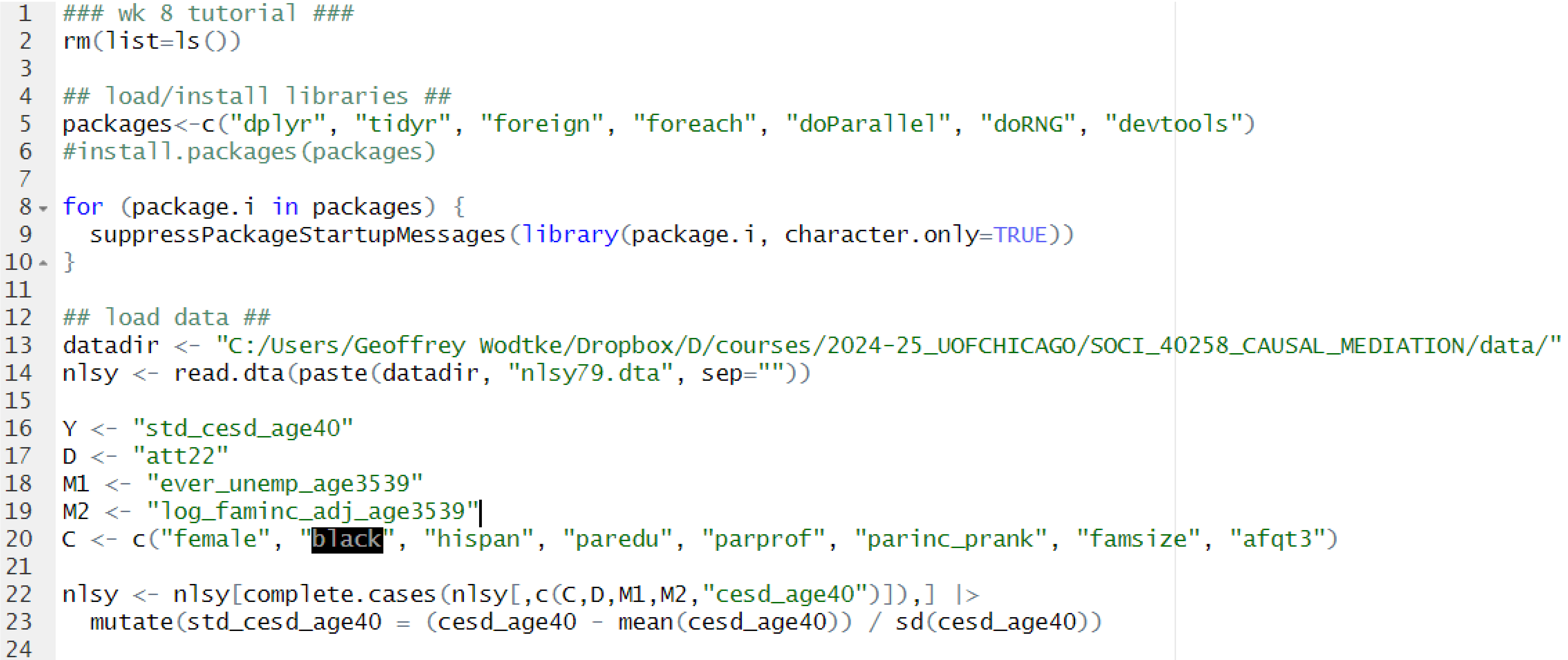
- Many studies have documented that going to college seems to reduce the likelihood of becoming depressed later in life—but how does this effect come about?

- One possibility is that a more advanced education reduces depression by increasing the labor market prospects of adults

- Do unemployment and income mediate the effect of college attendance on depression? If so, how? What is the respective contribution of different causal chains, or paths, to the total effect of education?

# Example: NLSY79

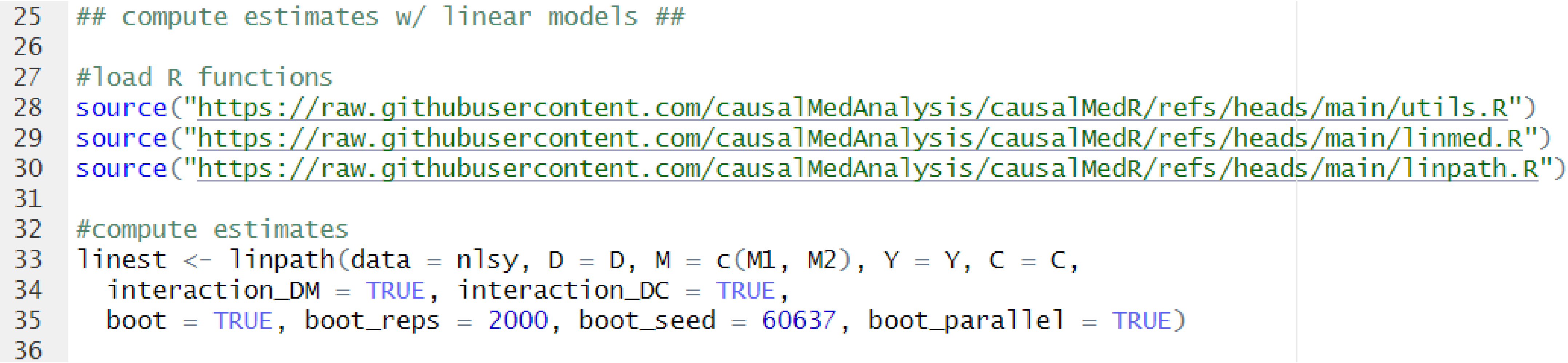
- Using linear models with interactions, compute estimates for the *PSEs* of education on depression operating through income and unemployment



```r
### wk 8 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/Geoffrey Wodtke/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"
M1 <- "ever_unemp_age3539"
M2 <- "log_faminc_adj_age3539"
C <- c("female", "black", "hispan", "paredu", "parprof", "parinc_prank", "famsize", "afqt3")

nlsy <- nlsy[complete.cases(nlsy[,c(C,D,M1,M2,"cesd_age40")]),] |>
  mutate(std_cesd_age40 = (cesd_age40 - mean(cesd_age40)) / sd(cesd_age40))

```

# Example: NLSY79

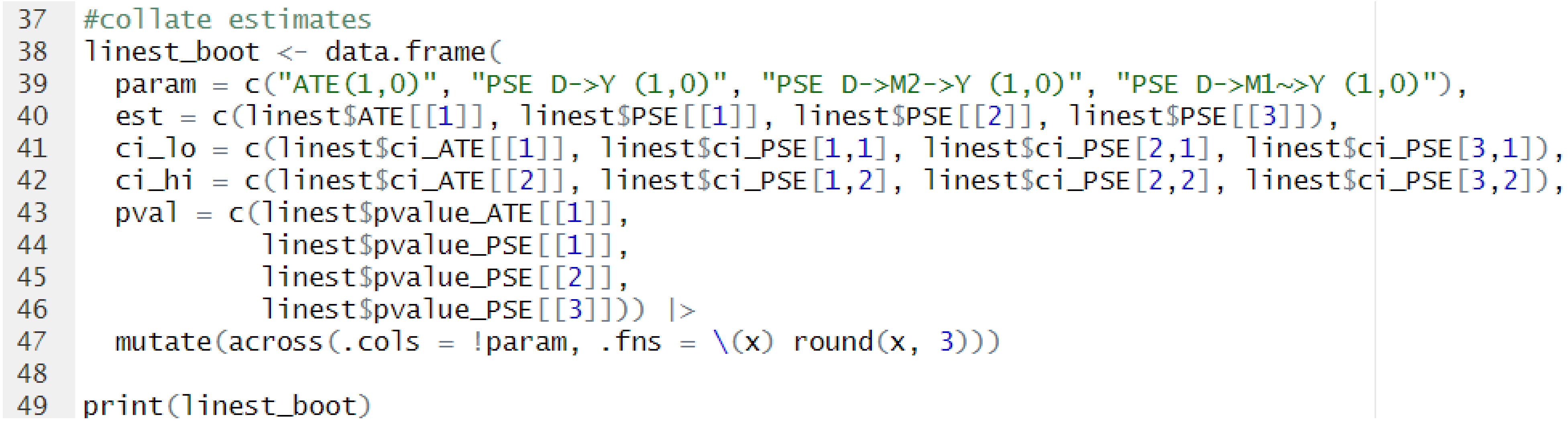
- Using linear models with interactions, compute estimates for the *PSEs* of education on depression operating through income and unemployment



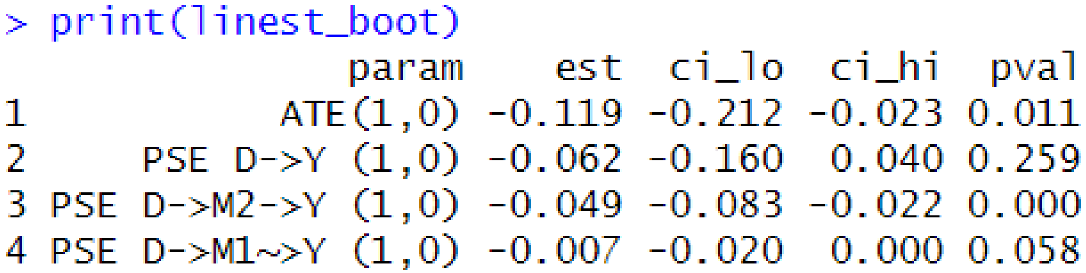
```r
## compute estimates w/ linear models ##

#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/linmed.R")
source("https://raw.githubusercontent.com/causalMedAnalysis/causalMedR/refs/heads/main/linpath.R")

#compute estimates
linest <- linpath(data = nlsy, D = D, M = c(M1, M2), Y = Y, C = C,
  interaction_DM = TRUE, interaction_DC = TRUE,
  boot = TRUE, boot_reps = 2000, boot_seed = 60637, boot_parallel = TRUE)

```

# Example: NLSY79

- Using linear models with interactions, compute estimates for the *PSEs* of education on depression operating through income and unemployment



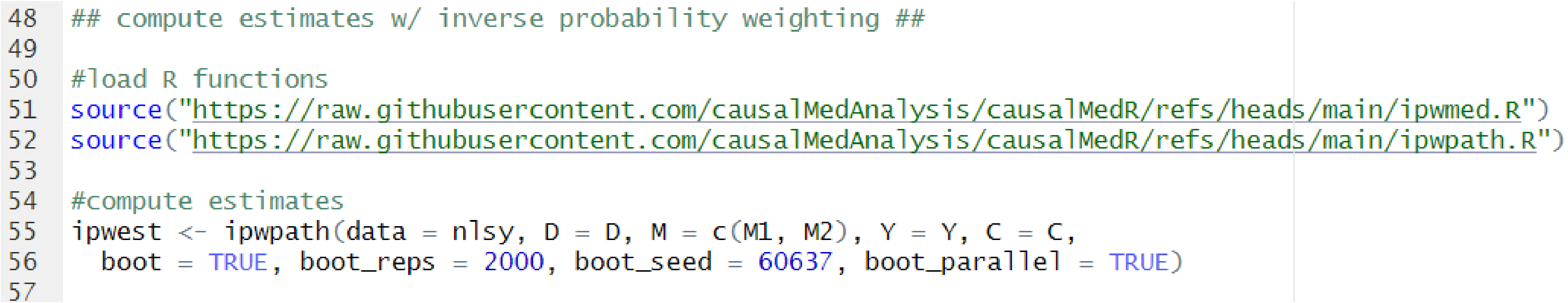
```r
#collate estimates
linest_boot <- data.frame(
  param = c("ATE(1,0)", "PSE D->Y (1,0)", "PSE D->M2->Y (1,0)", "PSE D->M1~>Y (1,0)"),
  est = c(linest$ATE[[1]], linest$PSE[[1]], linest$PSE[[2]], linest$PSE[[3]]),
  ci_lo = c(linest$ci_ATE[[1]], linest$ci_PSE[1,1], linest$ci_PSE[2,1], linest$ci_PSE[3,1]),
  ci_hi = c(linest$ci_ATE[[2]], linest$ci_PSE[1,2], linest$ci_PSE[2,2], linest$ci_PSE[3,2]),
  pval = c(linest$pvalue_ATE[[1]],
           linest$pvalue_PSE[[1]],
           linest$pvalue_PSE[[2]],
           linest$pvalue_PSE[[3]])) |>
  mutate(across(.cols = !param, .fns = \(x) round(x, 3)))

print(linest_boot)
```



```
> print(linest_boot)
               param    est  ci_lo  ci_hi  pval
1           ATE(1,0) -0.119 -0.212 -0.023 0.011
2     PSE D->Y (1,0) -0.062 -0.160  0.040 0.259
3 PSE D->M2->Y (1,0) -0.049 -0.083 -0.022 0.000
4 PSE D->M1~>Y (1,0) -0.007 -0.020  0.000 0.058
```

# Example: NLSY79

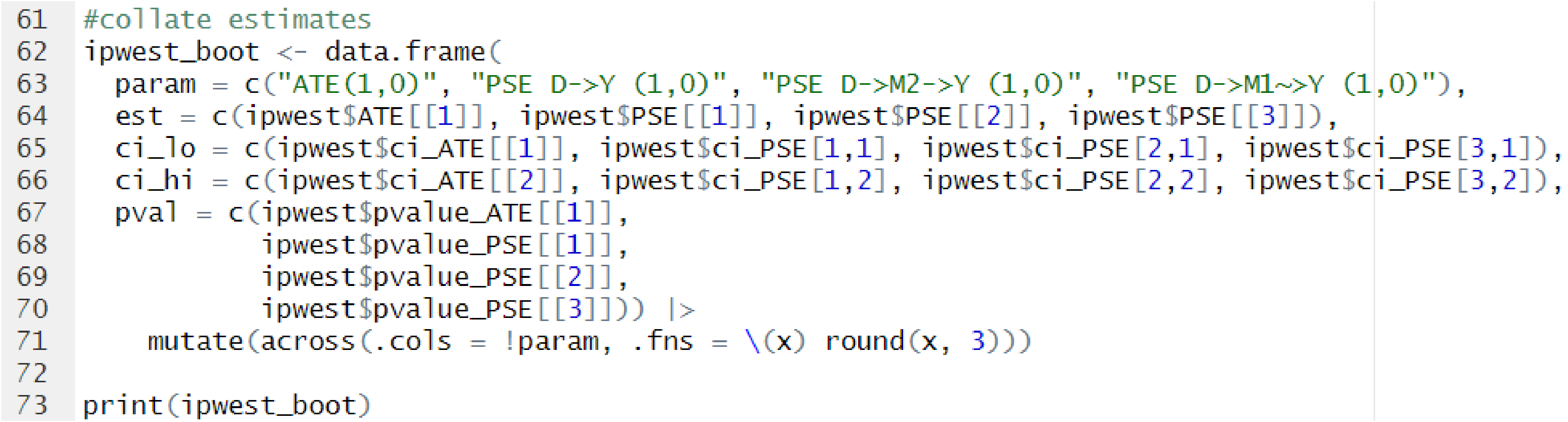
- Using inverse probability weighting, compute estimates for the *PSEs* of education on depression operating through income and unemployment



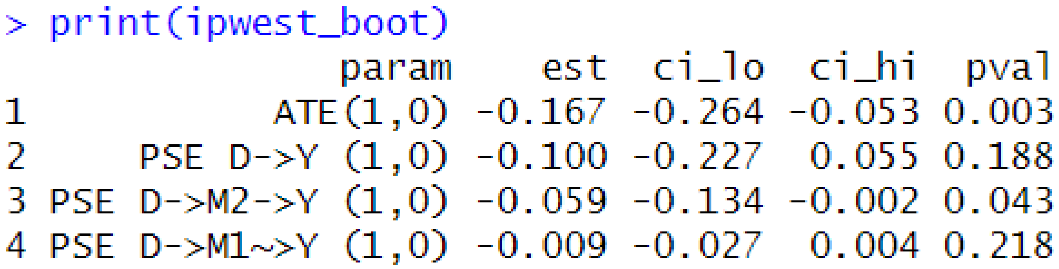
```r
## compute estimates w/ inverse probability weighting ##

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

#compute estimates
ipwest <- ipwpath(data = nlsy, D = D, M = c(M1, M2), Y = Y, C = C,
  boot = TRUE, boot_reps = 2000, boot_seed = 60637, boot_parallel = TRUE)

```

# Example: NLSY79

- Using inverse probability weighting, compute estimates for the *PSEs* of education on depression operating through income and unemployment



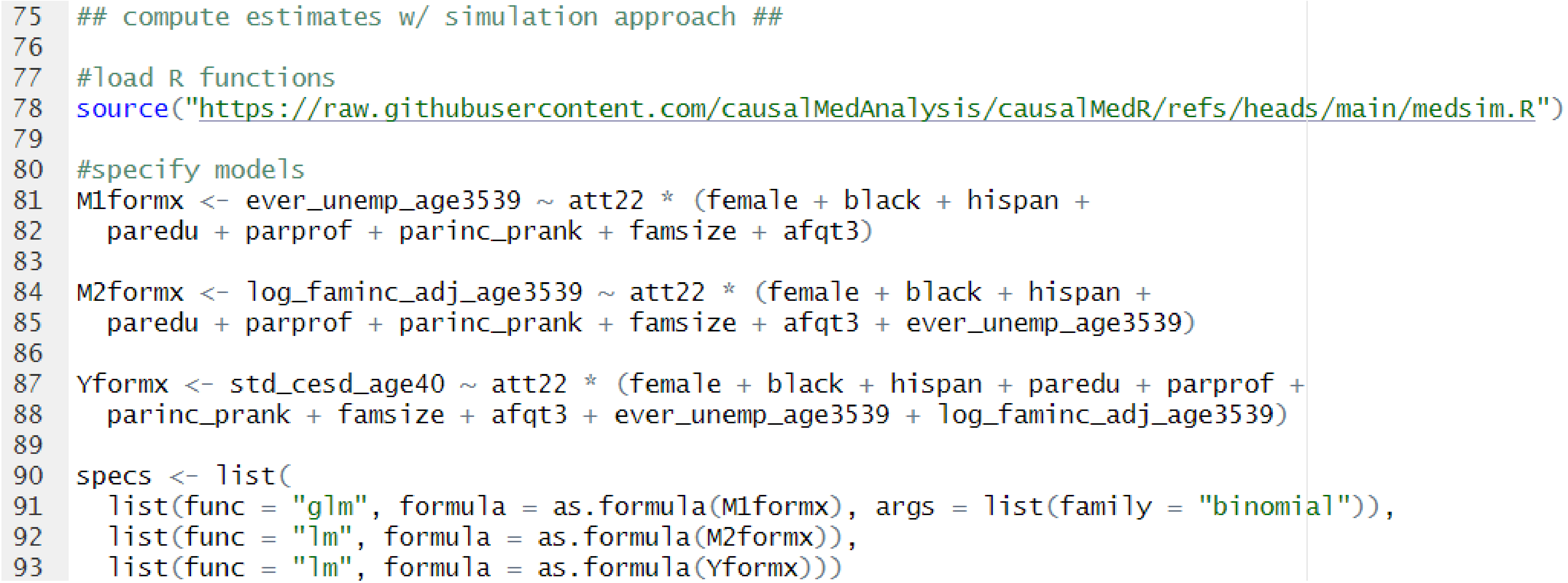
```r
#collate estimates
ipwest_boot <- data.frame(
  param = c("ATE(1,0)", "PSE D->Y (1,0)", "PSE D->M2->Y (1,0)", "PSE D->M1~>Y (1,0)"),
  est = c(ipwest$ATE[[1]], ipwest$PSE[[1]], ipwest$PSE[[2]], ipwest$PSE[[3]]),
  ci_lo = c(ipwest$ci_ATE[[1]], ipwest$ci_PSE[1,1], ipwest$ci_PSE[2,1], ipwest$ci_PSE[3,1]),
  ci_hi = c(ipwest$ci_ATE[[2]], ipwest$ci_PSE[1,2], ipwest$ci_PSE[2,2], ipwest$ci_PSE[3,2]),
  pval = c(ipwest$pvalue_ATE[[1]],
           ipwest$pvalue_PSE[[1]],
           ipwest$pvalue_PSE[[2]],
           ipwest$pvalue_PSE[[3]])) |>
    mutate(across(.cols = !param, .fns = \(x) round(x, 3)))

print(ipwest_boot)
```



```
> print(ipwest_boot)
                 param    est  ci_lo  ci_hi  pval
1             ATE(1,0) -0.167 -0.264 -0.053 0.003
2       PSE D->Y (1,0) -0.100 -0.227  0.055 0.188
3 PSE D->M2->Y (1,0) -0.059 -0.134 -0.002 0.043
4 PSE D->M1~>Y (1,0) -0.009 -0.027  0.004 0.218
```

# Example: NLSY79

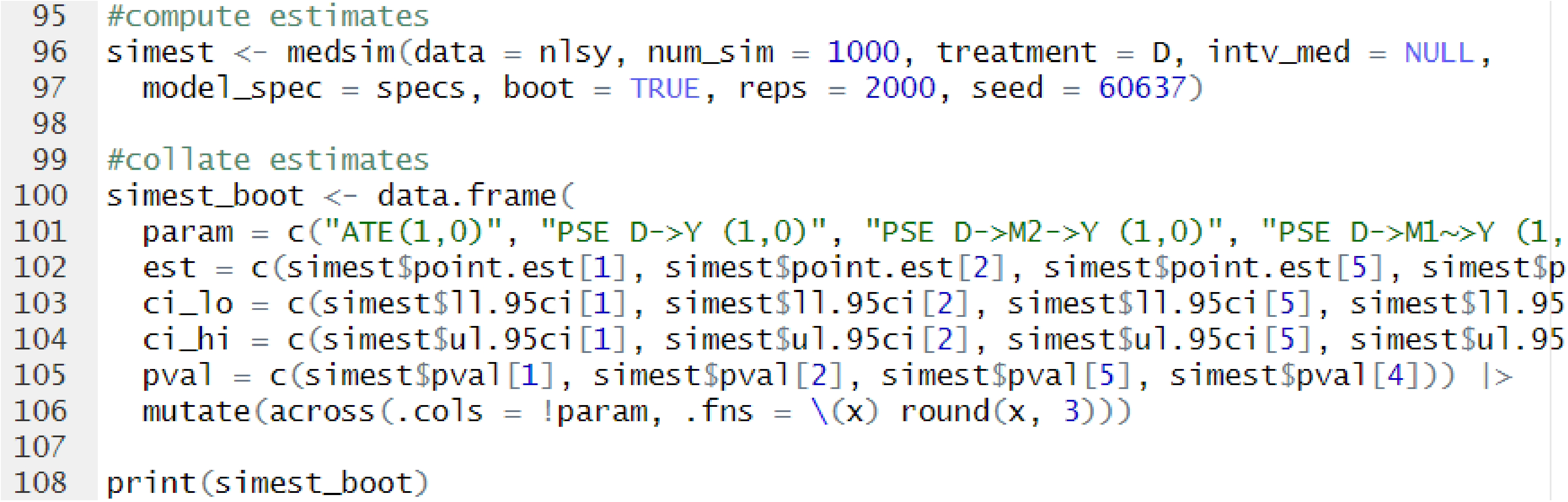
- Using the simulation approach, compute estimates for the $PSEs$ of education on depression operating through income and unemployment



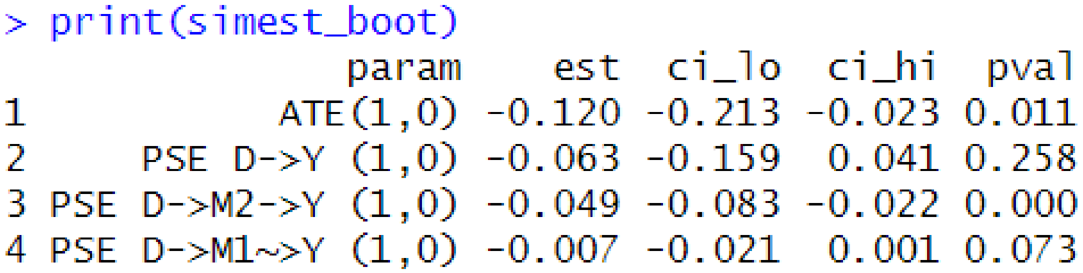
```
## compute estimates w/ simulation approach ##

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

#specify models
M1formx <- ever_unemp_age3539 ~ att22 * (female + black + hispan +
  paredu + parprof + parinc_prank + famsize + afqt3)

M2formx <- log_faminc_adj_age3539 ~ att22 * (female + black + hispan +
  paredu + parprof + parinc_prank + famsize + afqt3 + ever_unemp_age3539)

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

specs <- list(
  list(func = "glm", formula = as.formula(M1formx), args = list(family = "binomial")),
  list(func = "lm", formula = as.formula(M2formx)),
  list(func = "lm", formula = as.formula(Yformx)))
```

# Example: NLSY79

- Using the simulation approach, compute estimates for the *PSEs* of education on depression operating through income and unemployment



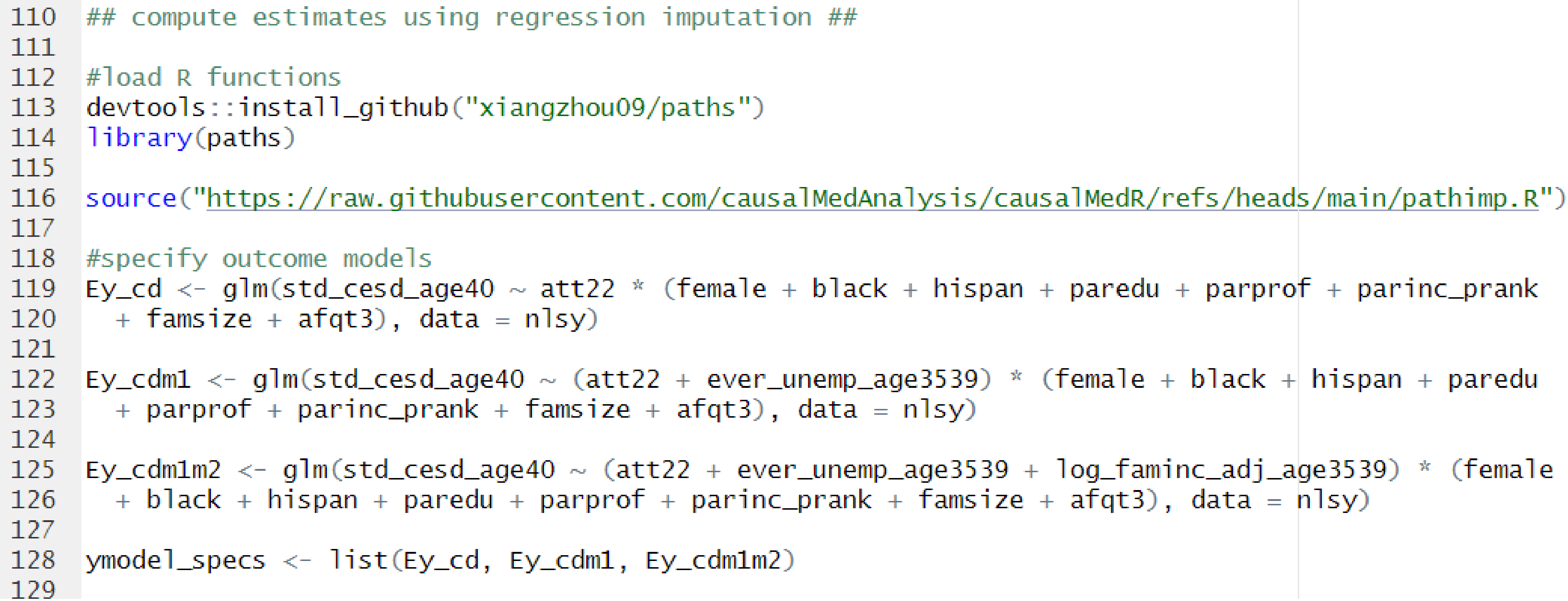
```r
#compute estimates
simest <- medsim(data = nlsy, num_sim = 1000, treatment = D, intv_med = NULL,
  model_spec = specs, boot = TRUE, reps = 2000, seed = 60637)

#collate estimates
simest_boot <- data.frame(
  param = c("ATE(1,0)", "PSE D->Y (1,0)", "PSE D->M2->Y (1,0)", "PSE D->M1~>Y (1,0)"),
  est = c(simest$point.est[1], simest$point.est[2], simest$point.est[5], simest$point.est[4]),
  ci_lo = c(simest$ll.95ci[1], simest$ll.95ci[2], simest$ll.95ci[5], simest$ll.95ci[4]),
  ci_hi = c(simest$ul.95ci[1], simest$ul.95ci[2], simest$ul.95ci[5], simest$ul.95ci[4]),
  pval = c(simest$pval[1], simest$pval[2], simest$pval[5], simest$pval[4])) |>
  mutate(across(.cols = !param, .fns = \(x) round(x, 3)))

print(simest_boot)
```



```
> print(simest_boot)
               param    est  ci_lo  ci_hi  pval
1           ATE(1,0) -0.120 -0.213 -0.023 0.011
2     PSE D->Y (1,0) -0.063 -0.159  0.041 0.258
3 PSE D->M2->Y (1,0) -0.049 -0.083 -0.022 0.000
4 PSE D->M1~>Y (1,0) -0.007 -0.021  0.001 0.073
```

# Example: NLSY79

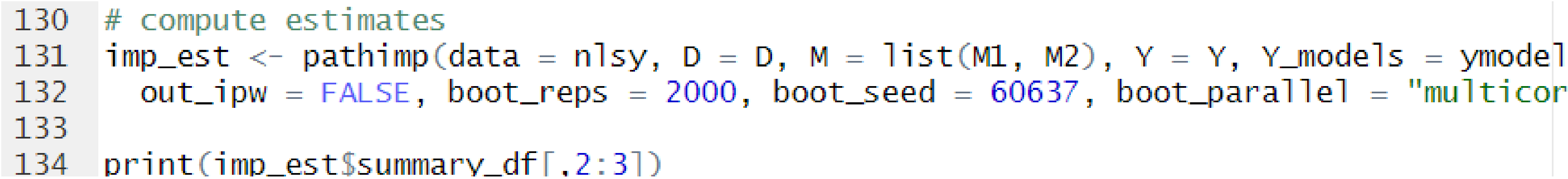
- Using regression imputation, compute estimates for the *PSEs* of education on depression operating through income and unemployment



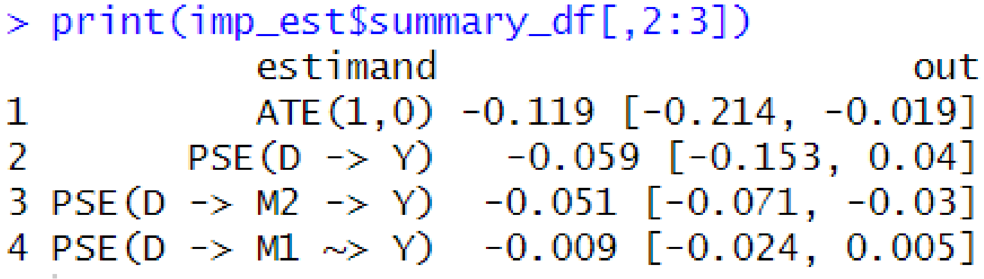
```
## compute estimates using regression imputation ##

#load R functions
devtools::install_github("xiangzhou09/paths")
library(paths)

source("https://raw.githubusercontent.com/causalMedAnalysis/causalMedR/refs/heads/main/pathimp.R")

#specify outcome models
Ey_cd <- glm(std_cesd_age40 ~ att22 * (female + black + hispan + paredu + parprof + parinc_prank
  + famsize + afqt3), data = nlsy)

Ey_cdm1 <- glm(std_cesd_age40 ~ (att22 + ever_unemp_age3539) * (female + black + hispan + paredu
  + parprof + parinc_prank + famsize + afqt3), data = nlsy)

Ey_cdm1m2 <- glm(std_cesd_age40 ~ (att22 + ever_unemp_age3539 + log_faminc_adj_age3539) * (female
  + black + hispan + paredu + parprof + parinc_prank + famsize + afqt3), data = nlsy)

ymodel_specs <- list(Ey_cd, Ey_cdm1, Ey_cdm1m2)

```

# Example: NLSY79

- Using regression imputation, compute estimates for the *PSEs* of education on depression operating through income and unemployment



```r
# compute estimates
imp_est <- pathimp(data = nlsy, D = D, M = list(M1, M2), Y = Y, Y_models = ymodel_specs,
  out_ipw = FALSE, boot_reps = 2000, boot_seed = 60637, boot_parallel = "multicore")

print(imp_est$summary_df[,2:3])
```



```
> print(imp_est$summary_df[,2:3])
            estimand                      out
1           ATE(1,0) -0.119 [-0.214, -0.019]
2        PSE(D -> Y)   -0.059 [-0.153, 0.04]
3 PSE(D -> M2 -> Y)   -0.051 [-0.071, -0.03]
4 PSE(D -> M1 ~> Y)   -0.009 [-0.024, 0.005]
```