

Exercise 3

Structural Causal Model, Causal Quantity, Identification

You're studying the effects of smoking marijuana on hypertension. For all subjects in your study, you've measured the hypertension outcome, the amount of time each patient has smoked marijuana, age, sex, other substance abuse habits, socioeconomic status, and race.

The target causal parameter is the difference in the counterfactual hypertension if all subjects had smoked marijuana daily for more than 6 months ($A = 1$) and the counterfactual hypertension if all subjects had smoked marijuana daily for no more than 6 months ($A = 0$)

$$\Psi^F(P_{U,X}) = E_{U,X}(Y_1) - E_{U,X}(Y_0)$$

where the counterfactual outcome Y_a is the hypertension outcome for an individual if, possibly contrary to fact, the individual had smoked marijuana according to $A = a$.

- 1) **Specify a structural causal model (SCM)** by answering parts (a)-(e). Use the formal notation defined in class. Recall that W denotes baseline (pre-exposure) covariates, A the exposure or treatment, and Y the outcome. Feel free to split your covariates into W_1, W_2, \dots as needed, but please define your random variables. If needed, use Z to denote a random variable, occurring after the exposure but before the outcome. The data collected should give you some idea of the covariates that could be included, but there may be additional ones. However, do not be too concerned if you are unsure of some of the relationships in your model.
 - a) What are the endogenous variables X ? What additional covariates besides those observed might be important? How would you incorporate them into the model?
 - b) Discuss your exogenous variables U . What factors might be included?
 - c) Specify your structural equations F . Do your structural equations make any assumptions about functional form?
 - d) Discuss any exclusion restrictions and any independence assumptions. In general, what are exclusion restrictions? In general, what are independence assumptions?

e) Draw at least one possible causal graph.

2) **Assess identifiability of $\Psi^F(P_{U,X})$** by answering parts (a)-(d). Recall that identifiability of the causal target parameter requires for each $P_{U,X}$ in \mathcal{M}^F (each $P_{U,X}$ compatible with the SCM) we have that $\Psi^F(P_{U,X}) = \Psi(P_0)$.

a) If not identified, under what assumptions would it be? Are some of these sets of additional assumptions more plausible than others? Are there additional measurements you could make so that the needed identifiability assumptions are more plausible?

b) What notation do we use to denote the original SCM, augmented with additional assumptions needed for identifiability?

c) Specify the target parameter of the observed data distribution (i.e. the statistical estimand).

d) What is the relevant positivity assumption? Are you concerned about violations of the positivity assumption in your study?