final project

Fanding Zhou

4/29/2022

Step 1: State scientific question.

(a) The causal question

What is the effect of a large-scale primary care redesign, the Comprehensive Primary Care Plus (CPC+) Initiative, on the monthly Medicare patient expenditures?

(b) The target population

Patients in different primary care practices in 18 regions.

Step 2: Specify a structural causal model (SCM).

(a) The endogenous variables V are variables that are meaningful for the scientific question and affected by other variables in this model. For this study, we define the following endogenous variables:

- $L_0 = \{V1, V2, V3, V4, V5\}_{t=2}$
- $L_1 = \{V1, V2, V3, V4, V5\}_{t=3}$
- V1: age, continuous variable supported on [0,100]
- V2: income groups, ordinal variable encoded into 0~14 integers.
- V3: sex, 0 for female, 1 for male, appromately equally distributed.
- V4: mean standardized Hierarchical Condition Category (HCC) score, approximately standard normal distribution.
- V5: race categories, white/black/all other, 70:20:10.
- V1-V5 are time variant patient-level covariates
- $X = \{X1, X2, X3, X4, X5, X6, X7, X8, X9\}$
- X1-X9 are time-invariant practice-level covariates.
- Z = Initiative of Comprehensive Primary Care Plus (CPC+) in year 3
- Y_0 = Monthly Medicare patient expenditures in year 2
- Y_1 = Monthly Medicare patient expenditures in year 3

Here $V = (C_0, C_1, T_0, T_1, Y_0, Y_1)$

(b) The exogenous U include all the unmeasured or unknown factors not included in V that impact the values that the V variables take.

The exogenous nodes are $U = (U_X, U_T, U_{L_0}, U_{L_1}, U_{Y_0}, U_{Y_1})$.

According to the guidance of ACIC competition, the DGPs will be free of unmeasured confounding. In other words, we can place some independence assumptions on the distribution of unmeasured factors P_U. To simplify, we assume the all the exogenous variables satisfy pairwise independence.

(c) This would suggest the following structural equations F:

$$X = f_X(U_X)$$
$$L_0 = f_{L_0}(U_{L_0})$$
$$Y_0 = f_{Y_0}(U_{Y_0}, L_0)$$
$$L_1 = f_{L_1}(U_{L_1}, L_0, Y_0)$$

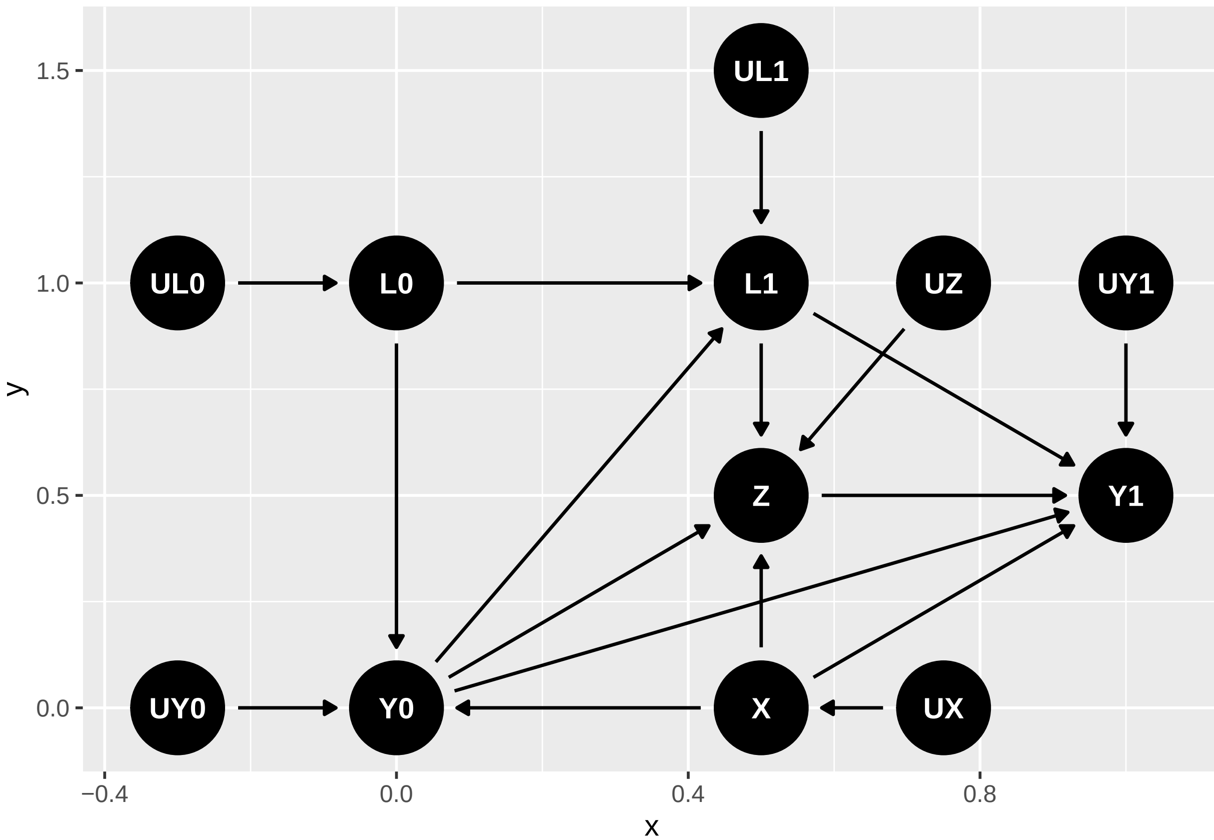
$$Z = f_Z(U_Z, L_1, Y_0, X)$$
$$Y_1 = f_{Y_1}(U_{Y_1}, L_1, T, Y_0, X)$$

(d) Exclusion restrictions.

An exclusion restriction is when a variable is not directly affected by another variable that precedes it in the structural causal model. For our model, we assume L_0 didn't directly affect T, but has some indirect effect through L_1 and Y_0. In addition, we assume practice level covariants X only have impact on Y_0, Y_1, T, but do not on those time-variant patient level covariants L_0 and L_1.

(f) Causal graph.

```
dag_m <- dagitty('dag {
    Y0 [pos="0,0"]
    L0 [pos="0,1"]
    Z [pos="0.5,0.5"]
    X [pos="0.5,0"]
    L1 [pos="0.5, 1"]
    Y1 [pos="1,0.5"]
    UL0[pos="-0.3,1"]
    UL1[pos="0.5,1.5"]
    UY0[pos="-0.3,0"]
    UZ[pos="0.75,1"]
    UX[pos="0.75,0"]
    UY1[pos="1,1"]
    UL0->L0
    UL1->L1
    UY0->Y0
    UZ->Z
    UX->X
    UY1->Y1
    L0->L1
    L0->Y0
    Y0->Z
    Y0->L1
    Y0->Y1
    L1->Y1
    L1->Z
    X->Y0
    X->Z
    X->Y1
    Z->Y1
    }')
ggdag(dag_m, layout = "circle")
```



Step 3. Specify the target parameter of the observed data distribution

(a) The counterfactuals of interest are Y(1) and Y(0) and they can be defined as

- $Y_1(1)$: Monthly Medicare patient expenditures if all patients had joined CPC+ in year 3.
- $Y_1(0)$: Monthly Medicare patient expenditures if all patients hadn't joined CPC+ in year 3.

(b) Our target parameter is the average causal effect:

$$\Psi_{U,V}(P_0) = \mathbb{E}_{U,V}[Y_1(1) - Y_1(0)]$$

This means the target causal parameter is the difference in the monthly Medicare patient expenditures for patients enrolled in CPC+ for 2 years and not enrolled in CPC+.

Step 4. Specify the observed data.

Suppose the data is from i.i.d sample from n = 329250 randomly sampled patients.

- Ti: whether patient i has enrolled in CPC+ in year 3
- Xi: practical level confounders collected for patient i
- $L_{0,i}$: patient level confounders collected for patient i in year 2
- $L_{1,i}$: patient level confounders collected for patient i in year 3
- $Y_{0,i}$: Monthly Medicare patient expenditures for patient i in year 2
- $Y_{1,i}$: Monthly Medicare patient expenditures for patient i in year 3

Further assume that the collected data are independent and identically distributed.

Step 5. Identify the targeted causal effect with the observed data.

(a) Conditional independent assumption

$$Y_{1,i}(0), Y_{1,i}(0) \perp\!\!\!\perp Z | X, L_1, Y_0$$

(b) By backdoor-path criterion, we can write our causal parameter with the observed data by G-computation:

$$\Psi(P_0) = \mathbb{E}[\mathbb{E}[Y_1|T=1, X, L_1, Y_0]] - \mathbb{E}[\mathbb{E}[Y_1|T=1, X, L_1, Y_0]]$$

Under our causal model and assumptions, average treatment effect equals to the observed difference in mean outcome within confounder strata, standardized to distribution of confounders.

(c) Relevant positivity assumption

Step 6. Estimation and Statistical Inference.

```
set.seed(252)
ObsData = read.table("filtered_patient_1.csv")
```

result:

```
ltmle.SL = readRDS("result1.rds")
summary(ltmle.SL)
```

```
## Estimator:  tmle
## Call:
## ltmle(data = ObsData, Anodes = "Z", Ynodes = "Y", abar = list(1,
##     0), SL.library = SL.library)
##
## Treatment Estimate:
##    Parameter Estimate:  1211.5
##     Estimated Std Err:  6.2942
##               p-value:  <2e-16
##     95% Conf Interval: (1199.1, 1223.8)
##
## Control Estimate:
##    Parameter Estimate:  1197.3
##     Estimated Std Err:  6.059
##               p-value:  <2e-16
##     95% Conf Interval: (1185.4, 1209.2)
##
## Additive Treatment Effect:
##    Parameter Estimate:  14.189
##     Estimated Std Err:  8.0474
##               p-value:  0.077874
##     95% Conf Interval: (-1.5838, 29.961)
```