Final Project: Childhood Bullying and Subsequent Drug Use

Shelley, Steph, Lizzy, Veronica

Spring 2019

##Structural Causal Model

Causal Question: What is the effect of having been bullied prior to age 12 on incidence of drug use in adolescence or adulthood?

Target Causal Parameter:

- Difference in the counterfactual probability of drug use if all kids were bullied prior to age 12, and the counterfactual probability of drug use if all kids were not bullied prior to age 12:

$$\psi^F(P_{U,X}) = P_{U,X}(Y_1 = 1) - P_{U,X}(Y_0 = 1) = E_{U,X}(Y_1) - E_{U,X}(Y_0)$$

where Y_a denotes the counterfactual outcome under an intervention to set bullying status A = a.

##Our Observed Data and Link to SCM

- Data are from the National Longitudinal Survey of Youth 1997
- A nationally representative cohort of youth age 12-16 (initial n=9000) in 1997, who were subsequently followed longitudinally.
- The target population is youth in the United States.

Data Description:

Key variables:

A: Bullying before the age of 12 (asked in 1997)

Y: Incident drug use ("cocaine or other hard drugs") after 1997

W: Race/ethnicity, age, sex, BMI, not living with both biological parents, mother's educational status (all Ws were measured at baseline)

Sample size: 7,703

We initially also considered: sexual orientation, learning/emotional disability, citizenship status

Marginal distribution of exposure and outcome:

Variable	No	Yes
Bullied < 12	6205	1498
Incident drug use	6373	1330

Identifiability

Estimand and Statistical Model

```
#just here as an example of a code chunk
mean(Y)
```

```
## [1] 0.17266
```

Estimation: Unadjusted ATE & SuperLearner

- The unadjusted ATE = mean(Y|A=1 - Y|A=0) = 0.05
- We use SuperLearner for prediction in all models.
  - Library: SL.glm, SL.glm.interaction, SL.glmnet, SL.bayesglm, SL.randomForest, SL.step, SL.mean, SL.loglinear
  - 5-fold cross-validation

Estimation: G-comp, IPTW, & TMLE

Estimator	ATE (95% CI)
G-computation	0.039
Stabilized IPTW	0.045
TMLE	0.044 (0.007, 0.08)

Estimation: SuperLearner convex combinations

Algorithm	A Risk	A Coefficient	Y Risk	Y Coefficient
glm	0.1549655	0	0.1408945	0.4628552
glm.interaction	0.155051	0.2086733	0.141501	0
glmnet	0.1549798	0	0.1409052	0
bayesglm	0.1549653	0	0.1408937	0
randomForest	0.1903391	0.4607251	0.1707125	0.2239811
step	0.1549472	0.267622	0.1408951	0.2476689
mean	0.1566939	0.0629796	0.1428772	0.0654947
loglinear	0.1549381	0	0.1409062	0

Estimation: SuperLearner performance

CV.SuperLearner

Algorithm	Avg Risk	SE
SuperLearner	0.1412062	0.0028693
Discrete SL	0.1407186	0.0027614
glm	0.1407131	0.0027616
glm.interaction	0.141278	0.0027675
glmnet	0.1407129	0.002763
bayesglm	0.1407127	0.0027616
randomForest	0.1678716	0.0041689
step	0.1407191	0.0027614
mean	0.1428802	0.0028201
loglinear	0.140726	0.0027623

Results