causal\_final

Siyan Chen

12/7/2019

# Problem 1

### a

### b.Employing traditional approaches to mediation analysis estimate total effect, direct effect, and difference method and product method indirect effect for a change in log and centered manganese levels from the 25th to the 75th percentile.

1 unit change in exposure-ln\_mn\_c:  
Total effect = -0.655  
direct effect==-0.352  
“difference method”-indirect effect==-0.655-(-0.352)=-0.303  
“product method”-indirect effect==-0.350\*0.864=-0.302

change in log and centered manganese levels from the 25th to the 75th percentile is 1.32. Therefore, it’s total effect is -0.865, direct effect is -0.465, indirect effect by “difference method” is -0.4 and indirect effect by “product method” is -0.4.

### c) Proceed calculating an estimate and 95% confidence interval for the interaction between the manganese exposure and birth length in a linear regression adjusted for the covariates.

The coefficient of interaction term is 0.277. The 95% confidence interval of thee coefficient is (0.042, 0.51)

### (d) If your goal is total effect decomposition, are the product and different method estimators valid in this context? Justify your answer.

It is not valid. In this context, we have mediator-exposure interaction term. Traditional methods cannot handel interaction issue. When interaction persents, difference method and product method will generate different results.

# Problem 2 possible unmeasured confounding

### (a) Consider the unmeasured confounder social economic status (SES). Draw a DAG representing the relationship between SES, manganese exposure, birth length, and neurodevelopment. Are natural direct and natural indirect effects identified? What about controlled direct effects? Justify your answer based on the assumptions required to identify each of these causal effects.

They cannot be identified.

To identify CDE, we need assumption of  
1) No unmeasured A->Y confounding given c  
2) No unmeasured M->Y confounding given c  
Based on the DAG, unmeasured SES is the common casue of manganese exposure and neurodevelopment outcome which d-connect A and Y. Unmeasured SES is also common cause of mediator birth length and neurodevelopment outcome which d-connect M and Y. Therefore, both assumptions are violated and we cannot identify CDE.

To identify natrual direct(NDE) and natural indirect effects(NIE), we need assumptions of  
3) No unmeasured A->M confounding given c  
4) No M->Y confounding that is caused by A  
Based on my DAG plot, unmeasured SES is the common cause of manganese exposure and birth length which d-connect A and M and violate the assumption 1 while assumption 2 is satisified. Therefore, we cannot identify NDE and NIE.

### (b) Hypothesize what could be the direction of the effect of SES on exposure, mediator and outcome (direction can be positive or negative). Given your hypothesis, can you conjecture the direction of confounding bias for direct and indirect effects? Justify your answer.

We can hypothesize that good SES has negative effect on exposure, has positive effect on mediator and positive effect on outcome.

Therefore, good SES can bias total effects to negative direction (protective effect) (+\*-=>-). Since good SES postively affect both mediator and outcome, the indirect effect (M to Y) bias to positive direction (causal effect). The direct effect bias to negative direction (protective effect)

# problem 3. Suppose now that we were able to collect data on other measures of fetal growth such gestational age.

### 3a.Draw the causal DAG in this setting.

### 3b.Are natural direct effect and natural indirect effect identified in the DAG you drew in 3a? Why or why not?

Yes, natural direct effect and natural indirect effect can be identified in the DAG, following assumptions are satisfied.  
1) No unmeasured A->Y confounding given c  
2) No unmeasured M->Y confounding given c  
3) No unmeasured A->M confounding given c  
4) No M->Y confounding that is caused by A

### 3c.How would you estimate controlled direct effects in the DAG you drew in 3a? Describe the estimation strategy in our own words (no calculations needed here.)

we can control mediator to one level and calculate the difference of expected value from A=1 to A=0.  
  
  
here we include fatal growth information in c  
  
And we follow same percedure to get CDE(M=m0)

# PART II

### 1). Look at the data. List all the variables of interest measured at each time point.

Time point 0: L1, L2, L3, A  
Time point 1: L1, L2, A  
Time point 2: Y

### 2) Write the complete causal DAG for this longitudinal observational study.

Let indicates covariate at time point j

### 3) Consider the following causal contrast:E[Y11-Y00],How would you interpret this causal contrast in your own words?

It is causal contrast between receiving treatment=1 at both time point 0 and time point 1 and receiving treatment=0 at both time point 0 and timpe point 1.

### 4) List the set of no unmeasured confounding assumptions that we need to satisfy in order to identify the causal contrast of interest.

Here, treatment at time0 (A0) can be considered as exposure and treatment at time1 (A1) can be considered as mediator.

No unmeasured exposure-outcome confounding given C   
No unmeasured mediator-outcome confounding given C   
No unmeasured exposure-mediator confounding given C   
No effect of exposure that confounds the mediator-outcome relationship

### 5) Provide an estimate and 95% confidence interval. Write the model specification and justify your model choice.

# beta\_0 mean  
mean(beta\_0)

## [1] -4.176892

# ci  
quantile(beta\_0, probs=c(0.025, 0.975))

## 2.5% 97.5%   
## -4.373470 -3.969055

# beta\_a0 mean  
mean(beta\_A0)

## [1] 0.02513366

# ci  
quantile(beta\_A0, probs=c(0.025, 0.975))

## 2.5% 97.5%   
## -0.1310451 0.1823873

# beta\_a1 mean  
mean(beta\_A1)

## [1] -0.3252439

# ci  
quantile(beta\_A1, probs=c(0.025, 0.975))

## 2.5% 97.5%   
## -0.4947432 -0.1772136

quantile(beta\_A0+beta\_A1,probs = c(0.5,0.025,0.975))

## 50% 2.5% 97.5%   
## -0.30078920 -0.58211111 -0.03302498

The fitted model is E(Y11-Y00)==-0.3 and it’s 95% confidence interval is (-0.582,-0.033)

### 6) Consider the following causal contrast: E[Y10-Y00].How would you interpret this causal contrast in your own words?

It is contrast comparing tratment level A0=1 to A0=0 setting A1=0.

### 7) List the set of no unmeasured confounding assumptions that we need to satisfy in order to identify the causal contrast of interest.

No unmeasured exposure-outcome confounding given C   
No unmeasured mediator-outcome confounding given C   
No unmeasured exposure-mediator confounding given C   
No effect of exposure that confounds the mediator-outcome relationship

### 8) Provide estimate and 95% confidence intervals. Write the model specification and justify your model of choice.

Fitted model is E[Y10-Y00]= and its 95% confidence interval is (-0.131 0.182)

### 9) Define natural direct and natural indirect effects in the context of this study. Are these causal effects identified? Justify your answer.

Natural direct effects: the average natural direct effect comparing treatment level at time point 0 (A0=1 to A0=0) when treament at time point 1 is set to natural condition.

Natutral indirect effects: the average natural indirect effect comparing treatment level at time point 1 (A1=1 to A1=0) when treament at time point 0 is set to natural condition.

Assuming no other unmasured confounding covariates, the following four assumption are followed. Therefore, causal effects can be identified.

1. No unmeasured A->Y confounding given c
2. No unmeasured M->Y confounding given c
3. No unmeasured A->M confounding given c
4. No M->Y confounding that is caused by A

df1 = read.csv("./data1\_final.csv")  
df2 = read.csv("./data2\_final.csv")  
head(df1)  
lm1 = lm(birthlength\_c~ln\_mn\_c+protein\_c+as.factor(female)+approxage, data=df1)  
summary(lm1)  
  
lm2 = lm(cognitive\_raw~ln\_mn\_c+protein\_c+as.factor(female)+approxage, data=df1)  
summary(lm2)  
  
lm3 = lm(cognitive\_raw~ln\_mn\_c+birthlength\_c+protein\_c+as.factor(female)+approxage, data=df1)  
summary(lm3)  
skimr::skim(df1)  
# 25th to the 75th percentile change in ln\_mn\_c  
# is from -0.61 to 0.021  
# change is 0.021-(-0.61)=0.631  
lm4 = lm(cognitive\_raw~ln\_mn\_c+birthlength\_c+ln\_mn\_c\*birthlength\_c+protein\_c+as.factor(female)+approxage, data=df1)  
summary(lm4)  
  
AC\_df = df2 %>% gather(key="Lt", value="value", c(L1,L2,L3,A)) %>%   
 arrange(id) %>%   
 mutate(Lt=str\_c(Lt,t0)) %>%   
 select(id, Lt, value) %>%  
 spread(key=Lt, value=value)   
Y\_df = df2 %>% select(id, L3,Y)   
wide\_df=merge(AC\_df, Y\_df) %>% select(-A2,-L22, -L12, -L31, -L32)  
  
set.seed(123)  
### bootstrap  
boots=1000  
beta\_0 = rep(NA, boots)  
beta\_A0 = rep(NA, boots)  
beta\_A1 = rep(NA, boots)  
nsize = nrow(wide\_df)  
for (i in 1:boots){  
 sample\_I = sample(1:nsize, size=nsize, replace = TRUE)  
 sample\_df = wide\_df[sample\_I,]  
 # time point 0  
 glm0 = glm(A0~L3, data=sample\_df, family = binomial)  
 p1 = predict(glm0, type="response")  
 w1=ifelse(sample\_df$A0==1, 1/p1, 1/(1-p1))  
 # time point 1  
 glm1 = glm(A1~A0+L3+L10+L20, data=sample\_df, family = binomial)  
 p2=predict(glm1, type="response")  
 w2=ifelse(sample\_df$A1==1, 1/p2, 1/(1-p2))  
 w=w1\*w2  
 sample\_df$w = w  
 model = svyglm(Y~A0+A1, family = gaussian(link = "identity"),design = svydesign(~1, weights=w, data=sample\_df))  
 beta\_0[i]= model$coef[1]  
 beta\_A0[i] = model$coef[2]  
 beta\_A1[i] = model$coef[3]  
}  
  
# beta\_0  
mean(beta\_0)  
quantile(beta\_0, probs=c(0.025, 0.975))  
  
# beta\_a0  
mean(beta\_A0)  
quantile(beta\_A0, probs=c(0.025, 0.975))  
  
# beta\_a1  
mean(beta\_A1)  
quantile(beta\_A1, probs=c(0.025, 0.975))  
  
quantile(beta\_A0+beta\_A1,probs = c(0.5,0.025,0.975))