

**Final Exam — Epi. 207a — 2000**  
Due: Wednesday, 8 November 2000, 3:30 p.m.

Do all questions. Report typos to Miguel or Jamie via email.

Throughout, ignore sampling variability. Work alone. Do not talk or share information with other students. You are required to come to the last class to grade exam on Wed. If you cannot come, contact Miguel.

**1.**

Table 1 contains data from an observational study of the direct effect of AZT ( $A_0$ ) on the mean level of HIV antigen ( $Y$ ) measured at time  $t_2$  controlling for the possible intermediate variable aerosolized pentamidine ( $A_1$ ). Here  $L_1 = 1$  if PCP develops by time  $t_1$ ,  $L_1 = 0$  otherwise.  $A_0 = 1$  if a subject takes AZT at time 0 and  $A_0 = 0$  otherwise.  $A_1 = 1$  if a subject takes aerosolized pentamidine at  $t_1$  and  $A_1 = 0$  otherwise. To help understand the last column, for any variable  $B$  note  $E(B | C = c)$  is by definition the mean (average) of the variable  $B$  among subjects whose value of the variable  $C$  is  $c$ . Thus, in row 1 of the table,  $E[Y | A_0, L_1, A_1] = 200$  is by definition  $E[Y | A_0 = 0, L_1 = 1, A_1 = 0]$ .

For the purpose of this exam, I am telling you that the following is valid: If you wish to check conditional independencies such as  $Y \perp\!\!\!\perp L_1 | A_1, A_0$  to see if  $L_1$  is a confounder, you may check whether  $E[Y | L_1 = 1, A_1, A_0] = E[Y | L_1 = 0, A_1, A_0]$ . Similarly, if, in doing g-estimation, you can check whether your guessed value for  $Y_{g=(0,0)}$  is independent of  $A_1$  given  $(L_1, A_0)$ , by checking whether  $E[Y_{g=(0,0)} | A_1 = 1, L_1, A_0] = E[Y_{g=(0,0)} | A_1 = 0, L_1, A_0]$ .

There are three scientists. Scientist 1 believes DAG 1 is the causal graph associated with these data. Scientist 2 believes DAG 2 is the causal graph. Scientist 3 believes DAG 3 is the causal graph. You are to answer the following questions **three (3)** times. First, you are to answer them as if you are Scientist 1. Second, you are to answer them as if you are Scientist 2. Third, you are to answer them as if you are Scientist 3. I would prefer if you answered all questions as Scientist 1, then all questions as Scientist 2, then all questions as Scientist 3, rather than answering Question 1 as all three scientists, then Question 2 as all three scientists, etc.

**1a.1.** Consider the complete statistical DAG  $A_0 \longrightarrow L_1 \longrightarrow A_1 \longrightarrow Y$ . Use the data in Table 1 to determine which, if any, of the arrows on this DAG can be deleted. Show your explicit numerical calculation justifying, for each arrow separately, either its deletion or retention.

SCIENTISTS 1, 2, and 3

To determine which, if any, arrows could be deleted you needed to check the following six conditional independence statements (one per arrow).

1.  $A_0 \perp\!\!\!\perp L_1$  for  $A_0 \longrightarrow L_1$
2.  $A_0 \perp\!\!\!\perp A_1 \mid L_1$  for  $A_0 \longrightarrow A_1$
3.  $A_0 \perp\!\!\!\perp Y \mid L_1, A_1$  for  $A_0 \longrightarrow Y$   
which implies  $E[Y \mid A_0 = 0, L_1, A_1] = E[Y \mid A_0 = 1, L_1, A_1]$
4.  $L_1 \perp\!\!\!\perp A_1 \mid A_0$  for  $L_1 \longrightarrow A_1$
5.  $L_1 \perp\!\!\!\perp Y \mid A_0, A_1$  for  $L_1 \longrightarrow Y$   
which implies  $E[Y \mid A_0, L_1 = 0, A_1] = E[Y \mid A_0, L_1 = 1, A_1]$
6.  $A_1 \perp\!\!\!\perp Y \mid A_0, L_1$  for  $A_1 \longrightarrow Y$   
which implies  $E[Y \mid A_0, L_1, A_1 = 1] = E[Y \mid A_0, L_1, A_1 = 0]$

If the statement is true, then the arrow can be removed. As only statement 2 is true, then only the arrow  $A_0 \longrightarrow A_1$  can be erased.

To check this, one has to show that  $\Pr(A_1 = 1 \mid A_0 = 0, L_1 = l_1) = \Pr(A_1 = 1 \mid A_0 = 1, L_1 = l_1)$ , for all values of  $l_1$ . This true for  $L_1 = 0$  (i.e.,  $0.25 = 0.25$ ) and for  $L_1 = 1$  (i.e.,  $0.75 = 0.75$ ).

**1a.2.** Use your causal DAG rather than the data in table 1 to determine which arrows on the complete statistical DAG can be removed.

(Note that even if your results on Questions 1a.1 and 1a.2 disagree, I want you to continue to assume your causal graph is correct for the remaining questions)

SCIENTISTS 1, 2 and 3

No arrows should be removed.

**1b.** Compute  $\Delta_0 = E[Y \mid A_0 = 1, A_1 = 0] - E[Y \mid A_0 = 0, A_1 = 0]$ . Based on your calculations in (1a) and your assumed causal graph, does  $\Delta_0$  have a causal interpretation as the direct effect of  $A_0$  when  $A_1$  is set to zero, i.e., as  $E[Y_{g=(1,0)}] - E[Y_{g=(0,0)}]$ . If yes, show why based on the rules you learned for confounding from the Pearl-Robins paper. If not, using these confounding rules, describe which arrows on your causal graph, were they missing, would allow such an interpretation. Repeat this question for  $\Delta_1 = E[Y \mid A_0 = 1, A_1 = 1] - E[Y \mid A_0 = 0, A_1 = 1]$ .

You had to calculate the weighted averages

$$E[Y \mid A_0 = 1, A_1 = 0] = \frac{1}{6000}(130 \times 3000 + 230 \times 3000) = 180$$

$$E[Y \mid A_0 = 0, A_1 = 0] = \frac{1}{8000}(50 \times 6000 + 200 \times 2000) = 87.5$$

and then the difference:  $\Delta_0 = 180 - 87.5 = 92.5$

Similarly for  $\Delta_1 = 124 - 182.5 = -58.5$ .

SCIENTIST 1

These differences have a causal interpretation

SCIENTISTS 2 and 3

These differences do not have a causal interpretation because  $L_1$  is a confounder. For these differences to have a causal interpretation, we would have to remove either the causal arrow from  $L_1$  to  $A_1$ , or the causal arrow from  $U_0$  to  $L_1$ , or the causal arrow from  $U_0$  to  $Y$ , or any combination of these.

**1c.** Is the fact that  $E[Y | A_0 = 1, L_1 = 0, A_1 = 1] - E[Y | A_0 = 0, L_1 = 0, A_1 = 1] = 250 - 70 = 180$  in Table 1 a valid reason to conclude that there is a direct causal effect of  $A_0$  on the mean of  $Y$  when  $A_1$  is set to one? If not, describe which arrows on your causal graph, were they missing, would allow such an interpretation.

SCIENTIST 1,2,3

It is not valid to assume that there is a direct causal effect of  $A_0$  on the mean of  $Y$  when  $A_1$  is set to zero. The reason is that  $L_1$  is a predictor of  $Y$  and is affected by earlier treatment. For this to have a causal interpretation, we would have to remove (1) the causal arrow from  $A_0$  to  $L_1$  and either the causal arrow from  $A_0$  to  $U_1$  or the causal arrow from  $U_1$  to  $L_1$  or (2) the causal arrow from  $U_0$  to  $L_1$ , or the causal arrow from  $U_0$  to  $Y$ , or both (1) and (2)..

**1d.** Could one validly estimate  $E[Y_{g=(0,0)}]$  by applying the g-computation algorithm to the statistical graph of Question 1a? Could one validly estimate  $E[Y_{g=(0,0)}]$  by applying the g-computation algorithm to the statistical complete graph  $A_0 \rightarrow A_1 \rightarrow Y$ ? Explain your answers.

SCIENTIST 1

No to first question because of path  $A_1 U_1 L_1 U_0 Y$ .. Yes to second question. The g-formula yields a valid estimate of  $E[Y_{g=(0,0)}]$  because  $L_1$  is not a confounder.

SCIENTIST 2

No to both questions. According to the Pearl-Robins sequential back-door theorem, there is no way to get a valid estimate of  $E[Y_{g=(0,0)}]$ , whether one uses data on  $L_1$  or not, i.e., there is intractable confounding.

SCIENTIST 3

Yes to the first question, No to the second one. The g-formula yields a valid estimate of  $E[Y_{g=(0,0)}]$  only if one uses data on  $L_1$  because  $L_1$  is a confounder.

## 2.

Consider the following potential structural nested models.

$$Y_{g=(0,0)} = Y_{g=(a_0, a_1)} - \beta_1 a_0 - \beta_2 a_1 - \beta_3 a_1 L_{1, g=(a_0)} \quad (2.1)$$

$$Y_{g=(0,0)} = Y_{g=(a_0,a_1)} - \beta_1 a_0 - \beta_2 a_1 \quad (2.2)$$

$$Y_{g=(0,0)} = Y_{g=(a_0,a_1)} - \beta_1 a_0 - \beta_2 a_1 - \beta_3 L_{1,g=(a_0)} \quad (2.3)$$

$$Y_{g=(0,0)} = Y_{g=(a_0,a_1)} - \beta_1 a_0 - \beta_2 a_1 - \beta_3 a_1 a_0 L_{1,g=(a_0)} \quad (2.4)$$

$$Y_{g=(0,0)} = Y_{g=(a_0,a_1)} - \beta_1 a_0 - \beta_2 a_1 - \beta_3 a_1 a_0 \quad (2.5)$$

(a): Are any of these not a structural nested model?

SCIENTISTS 1, 2, and 3

Model 2.3 is not a SNM because it is not consistent, i.e., for the values  $a_0 = 0$  and  $a_1 = 0$ , it states that  $Y_{g=(0,0)} = Y_{g=(0,0)} - \beta_3$ . This is not true for any set of data (unless  $\beta_3$  is assumed to be zero, which would imply that model 2.3 should be presented as  $Y_{g=(0,0)} = Y_{g=(a_0,a_1)} - \beta_1 a_0 - \beta_2 a_1$  to start with). We will not consider model 2.3 further.

(b): Assuming all  $\beta_2$  and  $\beta_3$  differ from zero, in which of the models does  $\beta_1 = 0$  correspond to the null hypothesis of no direct effect of  $A_0$  on each subject's  $Y$  controlling for  $A_1 = 0$ ?

SCIENTISTS 1, 2, and 3

Models 2.1, 2.2, 2.4, and 2.5 correspond to the null hypothesis of no direct effect of  $A_0$  on each subject's  $Y$  when  $A_1$  is set to 0 and  $\beta_1 = 0$ , because in all these models  $Y_{g=(0,0)} = Y_{g=(1,0)}$  when  $\beta_1 = 0$ .

(c): Answer the previous question with  $A_1 = 0$  replaced by  $A_1 = 1$

SCIENTISTS 1, 2, and 3

Model 2.2 corresponds to the null hypothesis of no direct effect of  $A_0$  on each subject's  $Y$  when  $A_1$  is set to 1 and  $\beta_1 = 0$ , because in this models  $Y_{g=(0,1)} = Y_{g=(1,1)}$  when  $\beta_1 = 0$ .

### 3.

Considering your assumed causal graph, is it possible to determine which, if any, of the models (2.1)-(2.5) is consistent with the data in our trial? If so, which is it and what values of  $(\beta_1, \beta_2, \beta_3)$  are implied by the data? Show your work. Do not use the G-computation algorithm formula or inverse probability weighting. Even if you cannot select a single model that you know is consistent with the data, can you eliminate any model because you know it is inconsistent with the data? Here, you may use all the results that you obtained in Question 1.

SCIENTIST 1

Based on the answer to question 1b and on your belief that there is no confounding, you know that

$$E [Y_{g=(0,0)}] = E [Y | A_0 = 0, A_1 = 0] = 87.5$$

$$E [Y_{g=(0,1)}] = E [Y | A_0 = 0, A_1 = 1] = 182.5$$

$$E [Y_{g=(1,0)}] = E [Y | A_0 = 1, A_1 = 0] = 180$$

$$E [Y_{g=(1,1)}] = E [Y | A_0 = 1, A_1 = 1] = 124$$

Then model 2.5 must be consistent with the data because it is a saturated model that allows one to estimate each of these 4 means without restrictions. The values of  $(\beta_1, \beta_2, \beta_3)$  are derived as follows:

$$E [Y_{g=(0,0)}] = E [Y_{g=(1,0)}] - \beta_1 \implies \beta_1 = 92.5$$

$$E [Y_{g=(0,0)}] = E [Y_{g=(0,1)}] - \beta_2 \implies \beta_2 = 95$$

$$E [Y_{g=(0,0)}] = E [Y_{g=(1,1)}] - 92.5 - 95 - \beta_3 \implies \beta_3 = -151$$

It is also intuitively clear that model 2.2 cannot be consistent with the data because of the restrictions it imposes (i.e., no interaction).

More formally, you can use g-estimation to study whether each model is consistent with the data. First, you calculate the counterfactual mean level of HIV antigen based on each model at each level of treatment history  $(A_0, L_1, A_1)$  as in table A.

Table A.

$A_0$	$L_1$	$A_1$	# subjects	Model 1	Model 2	Model 4	Model 5
0	1	0	2000	200	200	200	200
0	1	1	6000	$220 - \beta_2 - \beta_3$	$220 - \beta_2$	$220 - \beta_2$	$220 - \beta_2$
0	0	0	6000	50	50	50	50
0	0	1	2000	$70 - \beta_2$	$70 - \beta_2$	$70 - \beta_2$	$70 - \beta_2$
1	1	0	3000	$130 - \beta_1$	$130 - \beta_1$	$130 - \beta_1$	$130 - \beta_1$
1	1	1	9000	$110 - \beta_1 - \beta_2 - \beta_3$	$110 - \beta_1 - \beta_2$	$110 - \beta_1 - \beta_2 - \beta_3$	$110 - \beta_1 - \beta_2 - \beta_3$
1	0	0	3000	$230 - \beta_1$	$230 - \beta_1$	$230 - \beta_1$	$230 - \beta_1$
1	0	1	1000	$250 - \beta_1 - \beta_2$	$250 - \beta_1 - \beta_2$	$250 - \beta_1 - \beta_2$	$250 - \beta_1 - \beta_2 - \beta_3$

Then you have to find the model for which both  $Y_{g=(0,0)} \perp\!\!\!\perp A_1 | A_0$  and  $Y_{g=(0,0)} \perp\!\!\!\perp A_0$  hold. To check  $Y_{g=(0,0)} \perp\!\!\!\perp A_1 | A_0$ , you have to collapse table A over  $L_1$  to construct table B, and then compare  $E [Y_{g=(0,0)} | A_0, A_1 = 1]$  with  $E [Y_{g=(0,0)} | A_0, A_1 = 0]$  for all values of  $A_0$ . Because you believe there are no confounders, the model consistent with the data will show no differences between the pairs of counterfactual means. Note you cannot use Table A directly since  $Y_{g=(0,0)} \perp\!\!\!\perp A_1 | A_0, L_1$  is false.

Table B.

$A_0$	$A_1$	# subjects	Model 1	Model 2	Model 4	Model 5
0	0	8000	87.5	87.5	87.5	87.5
0	1	8000	$182.5 - \beta_2 - 0.75 \times \beta_3$	$182.5 - \beta_2$	$182.5 - \beta_2$	$182.5 - \beta_2$
1	0	6000	$180 - \beta_1$	$180 - \beta_1$	$180 - \beta_1$	$180 - \beta_1$
1	1	10000	$124 - \beta_1 - \beta_2 - 0.9 \times \beta_3$	$124 - \beta_1 - \beta_2$	$124 - \beta_1 - \beta_2 - 0.9 \times \beta_3$	$124 - \beta_1 - \beta_2 - \beta_3$

As expected, model 2.2 is misspecified. It simultaneously implies  $\beta_2 = 95$  (from  $87.5 = 182.5 - \beta_2$ ) and  $\beta_2 = 56$  (from  $180 - \beta_1 = 124 - \beta_1 - \beta_2$ ), which is a contradiction.

Also as expected, model 2.5 does not present any contradiction, and implies the following values:  $\beta_2 = 95, \beta_3 = -151$ , and no restrictions for  $\beta_1$ . Model 2.4 implies the values  $\beta_2 = 95, \beta_3 = -167.78$ , and no restrictions for  $\beta_1$ . Model 2.1 implies the values  $\beta_2 = 850, \beta_3 = -1006.67$ , and no restrictions for  $\beta_1$ .

Now you have to collapse over  $A_1$  and  $L_1$  (table C) to check  $Y_{g=(0,0)} \perp\!\!\!\perp A_0$  by comparing  $E[Y_{g=(0,0)} | A_0 = 1]$  and  $E[Y_{g=(0,0)} | A_0 = 0]$ .

Table C.

$A_0$	# subjects	Model 1	Model 2	Model 4	Model 5
0	16000	$135 - 0.5 \times \beta_2 - 0.375 \times \beta_3$	$135 - 0.5 \times \beta_2$	$135 - 0.5 \times \beta_2$	$135 - 0.5 \times \beta_2$
1	16000	$145 - \beta_1 - \frac{10}{16}\beta_2 - \frac{9}{16}\beta_3$	$145 - \beta_1 - \frac{10}{16}\beta_2$	$145 - \beta_1 - \frac{10}{16}\beta_2 - \frac{9}{16}\beta_3$	$145 - \beta_1 - \frac{10}{16}\beta_2 - \frac{10}{16}\beta_3$

Table C implies  $\beta_1$  is equal to 92.5 in all three models 2.1, 2.4, and 2.5. Thus models 2.1, 2.4, and 2.5 (but not 2.2) are consistent with the data.

### SCIENTIST 2

You only know that  $Y_{g=(0,0)} \perp\!\!\!\perp A_0$ . This is not sufficient to exclude any model as inconsistent nor to estimate the values of the parameters  $(\beta_1, \beta_2, \beta_3)$  or  $(\beta_1, \beta_2)$  of any model consistent with the data.

### SCIENTIST 3

You need to use g-estimation based on your belief that there are no unmeasured confounders given  $L_1$ , i.e., both  $Y_{g=(0,0)} \perp\!\!\!\perp A_1 | A_0, L_1$  and  $Y_{g=(0,0)} \perp\!\!\!\perp A_0$  hold.

To check which models are consistent with  $Y_{g=(0,0)} \perp\!\!\!\perp A_1 | A_0, L_1$  you have to use table A. and compare  $E[Y_{g=(0,0)} | A_0, L_1, A_1 = 1]$  with  $E[Y_{g=(0,0)} | A_0, L_1, A_1 = 0]$  for all values of  $A_0$  and  $L_1$ . Because there are no unmeasured confounders, models consistent with the data will show no differences between the pairs of counterfactual means. Only model 2.4 meets this requirement.

Equating the third and fourth rows of table A (i.e.,  $50 = 70 - \beta_2$ ), we get  $\beta_2 = 20$  for all models. Model 2.1 cannot be correct as it simultaneously implies  $\beta_3 = 0$  (from  $200 = 220 - \beta_2 - \beta_3$ ) and  $\beta_3 = -40$  (from  $130 - \beta_1 = 110 - \beta_1 - \beta_2 - \beta_3$ ), which is a contradiction. Similarly, model 2.2 implies  $\beta_2 = 20$  and  $\beta_2 = -20$ , and model 2.5 implies  $\beta_3 = 0$  and  $\beta_3 = -40$ .

Let us now check model 2.4. Equating rows 1 and 2, or 3 and 4, or 7 and 8 we get  $\beta_2 = 20$ . From rows 5 and 6 (i.e.,  $130 - \beta_1 = 110 - \beta_1 - \beta_2 - \beta_3$ ), it is deduced that  $\beta_3 = -40$ . There is no contradiction here ( $\beta_1$  may take any value for now).

Then you have to check  $Y_{g=(0,0)} \perp\!\!\!\perp A_0$ , i.e., whether  $E[Y_{g=(0,0)} | A_0 = 1] = E[Y_{g=(0,0)} | A_0 = 0]$  for model 2.4. This can be used to calculate  $\beta_1$ . From table 3, we see that, in order for this equality to hold,  $\beta_1$  must be equal to 30. Model 2.4 is consistent with the data.

#### 4.

If you were able to estimate the values of  $(\beta_1, \beta_2, \beta_3)$  in Question 3, then, based on the data and your choice of model in Question 3, compute the mean of  $Y_{g=(0,0)}$  [i.e.,  $E(Y_{g=(0,0)})$ ] for the study population.

SCIENTIST 1

$E(Y_{g=(0,0)}) = 87.5$  for all rows of model 2.5 in table C and hence  $E(Y_{g=(0,0)}) = 87.5$  for the study population. Same if you use models 2.1 or 2.4, or a weighted average of the values  $E(Y_{g=(0,0)})$  in each column corresponding to models 2.1, 2.4, or 2.5 in table A. Note row  $70 - \beta_2$  has negative value in table 1]

SCIENTIST 2

Impossible to compute.

SCIENTIST 3

You first had to use model 2.4 to calculate:  $E(Y_{g=(0,0)})$  for each row of table A, and then compute the weighted average

$$\frac{1}{32000}(200 \times 2000 + 200 \times 6000 + 50 \times 6000 + 50 \times 2000 + 100 \times 3000 + 100 \times 9000 + 200 \times 3000 + 200 \times 1000) = 125$$

More easily you could use table 3.

Considering your answers to Questions 2 and 3 only, can you determine whether there is a direct effect of  $A_0$  on the mean of  $Y$  controlling for  $A_1$  when  $A_1 = 1$ ? Compute if possible  $E[Y_{g=(1,1)}] - E[Y_{g=(0,1)}]$ .

How about when  $A_1 = 0$ ? Compute, if possible,  $E[Y_{g=(1,0)}] - E[Y_{g=(0,0)}]$ .

SCIENTIST 1

Yes to both questions. Using model 2.5,

$$E[Y_{g=(1,1)}] - E[Y_{g=(0,1)}] = (E[Y_{g=(0,0)}] + \beta_1 + \beta_2 + \beta_3) - (E[Y_{g=(0,0)}] + \beta_2) = \beta_1 + \beta_3 = -58.5$$

and

$$E[Y_{g=(1,0)}] - E[Y_{g=(0,0)}] = (E[Y_{g=(0,0)}] + \beta_1) - E[Y_{g=(0,0)}] = \beta_1 = 92.5$$

SCIENTIST 2

Direct effects cannot be computed.

SCIENTIST 3

No to the first question. Yes to the second one. Using model 2.4,  $E[Y_{g=(1,1)}] - E[Y_{g=(0,1)}] = (E[Y_{g=(0,0)}] + \beta_1 + \beta_2 + \beta_3 L_{1,g=(1)}) - (E[Y_{g=(0,0)}] + \beta_2) = \beta_1 + \beta_3 L_{1,g=(1)} = 30 - 40 L_{1,g=(1)}$ . This difference depends on the values of  $L_{1,g=(1)}$  which cannot be calculated with model 2.4. Therefore, it is not possible to determine the direct effect of  $A_0$  on the mean of  $Y$  controlling for  $A_1 = 1$ .

On the other hand,

$$E[Y_{g=(1,0)}] - E[Y_{g=(0,0)}] = (E[Y_{g=(0,0)}] + \beta_1) - E[Y_{g=(0,0)}] = \beta_1 = 30$$

**5.**

Is it possible to compute the effect of  $A_0$  on  $L_1$ ? If so, calculate  $E [L_{1,g=(a_0=0)}]$  and  $E [L_{1,g=(a_0=1)}]$  from the data.

SCIENTISTS 1, 2, and 3

Yes.

$$E [L_{1,g=(a_0=0)}] = \Pr[L_1 (a_0 = 0) = 1] = \Pr[L_1 = 1 | A_0 = 0] = \frac{8000}{16000} = 0.50$$

$$E [L_{1,g=(a_0=1)}] = \Pr[L_1 (a_0 = 1) = 1] = \Pr[L_1 = 1 | A_0 = 1] = \frac{12000}{16000} = 0.75$$

$$\text{The effect is } E [L_{1,g=(a_0=1)}] - E [L_{1,g=(a_0=0)}] = 0.25$$

**6.**

Considering only your answers to questions 2, 3, and 6, can you compute  $E [Y_{g=(1,k)}] - E [Y_{g=(0,k)}]$  for  $k = 0$  and  $k = 1$ . If so, compute it.

SCIENTISTS 1 and 2

Same answer as in question 5.

SCIENTIST 3

For  $k = 0$  same answer as in question 5. For  $k = 1$ , use  $E [L_{1,g=(a_0=1)}] = 0.75$  from question. 6 to get  $E [Y_{g=(1,1)}] - E [Y_{g=(0,1)}] = 30 - 40L_{1,g=(1)} = 0$ . There is no direct effect of  $A_0$  on the mean of  $Y$  controlling for  $A_1 = 1$ .

Can you compute  $E [Y_g]$  from these answers where  $g$  is the dynamic regime “take  $A_0$ . Then take  $A_1$  if  $L_1 = 1$  but do not take  $A_1$  if  $L_1 = 0$ ?” If so, compute it.

SCIENTIST 1: No depends on the correct model

SCIENTIST 2: No depends on the correct model

SCIENTIST 3: Yes

$$Y_g = Y_{g=(1,1)} \text{ if } L_{1,g=(1)} = 1 \text{ i.e. } Y_{g=(0,0)} + \beta_1 + \beta_2 + \beta_3$$

$$Y_g = Y_{g=(1,0)} \text{ if } L_{1,g=(1)} = 0 \text{ i.e. } Y_{g=(0,0)} + \beta_1$$

$$E [Y_g] = E [Y_{g=(0,0)}] + \beta_1 + (\beta_2 + \beta_3) E [L_{1,g=(1)} = 1]$$

**7.**

If possible, use the G-computation algorithm applied to the appropriate statistical graph, to compute  $E [Y_{g=(a_0,a_1)}]$  for each of the four joint levels of  $(a_0, a_1)$ .

SCIENTIST 1

$$E[Y_{g=(0,0)}] = E[Y | A_0 = 0, A_1 = 0] = 87.5$$

$$E[Y_{g=(0,1)}] = E[Y | A_0 = 0, A_1 = 1] = 182.5$$

$$E[Y_{g=(1,0)}] = E[Y | A_0 = 1, A_1 = 0] = 180$$

$$E[Y_{g=(1,1)}] = E[Y | A_0 = 1, A_1 = 1] = 124$$

SCIENTIST 2

Can't apply the g-computation algorithm.

SCIENTIST 3

You had to calculate  $E[Y_{g=(a_0, a_1)}] = E[Y(a_0, a_1)] = \sum_{l_1} E[Y | A_0 = a_0, L_1 = l_1, A_1 = a_1] \Pr[L_1 = l_1 | A_0 = a_0]$  for each combination of treatment, as it is shown below.

$A_0$	$A_1$	g-formula estimate
0	0	$200 \times \frac{8000}{16000} + 50 \times \frac{8000}{16000} = 125$
0	1	145
1	0	155
1	1	145

The answer  $E[Y_{g=(0,0)}] = 125$  obtained through the g-formula is the same as the one obtained by using our structural nested model 2.4 in question 4.

If possible, use the G-computation algorithm applied to the appropriate statistical graph, to compute  $E[Y_g]$  for the dynamic regime  $g$  of Question 7.

SCIENTIST 1: Not possible

SCIENTIST 2: Not possible

SCIENTIST 3:

$$\begin{aligned} E[Y_g] &= E[Y | A_0 = 1, L_1 = 1, A_1 = 1] \times \Pr[L_1 = 1 | A_0 = 1] + E[Y | A_0 = 1, L_1 = 0, A_1 = 0] \times \Pr[L_1 = 0 | A_0 = 1] \\ &= 110 \times \frac{12}{16} + 230 \times \frac{4}{16} = 140 \end{aligned}$$

8. .

9a. If possible, determine whether there is a direct effect of  $A_0$  on the mean of  $Y$  controlling for  $A_1$  when  $A_1 = 1$  using the G-computation algorithm results in Question 8.

SCIENTIST 1

$$E[Y_{g=(1,1)}] - E[Y_{g=(0,1)}] = 124 - 182.5 = -58.5$$

SCIENTIST 2

Not possible

SCIENTIST 3

$$E[Y_{g=(1,1)}] - E[Y_{g=(0,1)}] = 145 - 145 = 0$$

9b. Repeat (9a) except with  $A_1 = 0$ .

SCIENTIST 1

$$E[Y_{g=(1,0)}] - E[Y_{g=(0,0)}] = 180 - 87.5 = 92.5$$

SCIENTIST 2

Not possible

SCIENTIST 3

$$E[Y_{g=(1,0)}] - E[Y_{g=(0,0)}] = 155 - 125 = 30$$

Do your answers to this question agree quantitatively with those obtained in Question 7?

SCIENTISTS 1, 2, and 3

Yes

9.

Create a pseudo-population from the data in Table 1 using the stabilized weights  $SW = \frac{f(A_0)f(A_1|A_0)}{f(A_0)f(A_1|A_0, L_1)}$ . Show the pseudo-population data in a Table similar to Table 1. (You may have fractional people, if that is required.)

SCIENTIST 1,2,3

$A_0$	$L_1$	$A_1$	# subjects	$E[Y A_0, L_1, A_1]$	$f(A_1 A_0)$	$f(A_1 L_1, A_0)$	$SW$	Pseudo-pop.
0	1	0	2000	200	0.50	0.25	2	4000
0	1	1	6000	220	0.50	0.75	$\frac{2}{3}$	4000
0	0	0	6000	50	0.50	0.75	$\frac{2}{3}$	4000
0	0	1	2000	70	0.50	0.25	2	4000
1	1	0	3000	130	0.375	0.25	1.5	4500
1	1	1	9000	110	0.625	0.75	$\frac{5}{6}$	7500
1	0	0	3000	230	0.375	0.75	0.5	1500
1	0	1	1000	250	0.625	0.25	2.5	2500

10.

Answer Questions 1a.1 again, but this time treating the pseudo-population as the population to which the question refers. Is your answer using the pseudo-population data the same or different from your answer obtained using the actual population data? Explain any difference.

SCIENTIST 1,2,3

The arrow  $L_1 \rightarrow A_1$  does not exist in the pseudo-population. The arrow  $A_0 \rightarrow A_1$  is now present.

## 11.

Consider the marginal structural model

$$E[Y_{g=(a_0, a_1)}] = \theta_0 + \theta_1 a_0 + \theta_2 a_1 + \theta_3 a_0 a_1.$$

Can you use your pseudo-population data to calculate  $\theta_0$ ,  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$ ? If so, do the calculation in the easiest possible way.

SCIENTIST 1

No confounding induced.

SCIENTIST 2

No

SCIENTIST 3

Collapse the pseudo population over L and do a standard analysis.

$A_0$	$A_1$	Pseudo-pop.	$E[Y A_0 = a_0, A_1 = a_1]$	
0	0	8000	125	$= \theta_0$
0	1	8000	145	$= \theta_0 + \theta_2$
1	0	6000	155	$= \theta_0 + \theta_1$
1	1	10000	145	$= \theta_0 + \theta_1 + \theta_2 + \theta_3$

From the above table,  $\theta_0 = 125, \theta_1 = 30, \theta_2 = 20, \theta_3 = -30$ .

Can you use the pseudo-population data to calculate  $E[Y_g]$  for the dynamic regime of Question 7? If so how would you do it?

Apply the G-computation algorithm to the pseudopopulation data based on the graph is Q1a that uses  $L_1$ .

If you were not able to use the pseudo-population data to calculate the  $\theta'$ s, can you calculate them using another method? If so, do the calculation and explain.

For scientist 1 do calculation above for scientist 3 but applied to the real rather than the pseudo population.

### 12a.

The hypothesis that  $A_0$  has no direct effect on  $Y$  when  $A_1$  is set to 1 implies what restrictions on the values of  $(\theta_1, \theta_2, \theta_3)$ ?

SCIENTISTS 1, 2, and 3: The hypothesis is  $E[Y_{g=(1,1)}] = E[Y_{g=(0,1)}]$ , which would imply  $\theta_0 + \theta_1 + \theta_2 + \theta_3 = \theta_0 + \theta_2$ . This would be true if  $\theta_1 + \theta_3 = 0$  only.

### 12b.

The hypothesis that  $A_0$  has no direct effect on  $Y$  when  $A_1$  is set to 0 implies what restrictions on the values of  $(\theta_1, \theta_2, \theta_3)$ ?

SCIENTISTS 1, 2, and 3: The hypothesis is  $E[Y_{g=(1,0)}] = E[Y_{g=(0,0)}]$ , which would imply  $\theta_0 + \theta_1 = \theta_0$ . This would be true if  $\theta_1 = 0$  only.

## 12. .

Consider the association model  $E[Y | A_0 = a_0, A_1 = a_1] = \gamma_0 + \gamma_1 a_0 + \gamma_2 a_1 + \gamma_3 a_0 a_1$ . Use the data in Table 1 to calculate  $\gamma_0, \gamma_1, \gamma_2$ , and  $\gamma_3$ . Assuming you are able to estimate the  $\theta$ 's in Question 12, do the  $\gamma$ 's agree with the corresponding  $\theta$ 's in Question 12? Why or why not?

Using same approach as in last question  $\gamma_0 = 87.5, \gamma_1 = 92.5, \gamma_2 = 95$ , and  $\gamma_3 = -151$  for all three scientists.

SCIENTIST 1

The values for  $\gamma_0, \gamma_1, \gamma_2$ , and  $\gamma_3$  are the same as  $\theta_0, \theta_1, \theta_2$ , and  $\theta_3$  because the pseudopopulation is the same as the actual population (i.e., you assume there is no confounding).

SCIENTIST 2

No  $\theta$ 's estimated.

SCIENTIST 3

The  $\gamma$ 's and the  $\theta$ 's differ because there is confounding.

## 13. .

Describe precisely how you would use the least squares package in SAS to compute  $\gamma_0, \gamma_1, \gamma_2$ , and  $\gamma_3$  in Question 13 if you had the actual raw data on each study subject in Table 1. That is, write in detail the necessary SAS code.

SCIENTISTS 1, 2, and 3

```
proc reg;
model y= a0 a1 a0a1;
* a0a1 created in a previous data step as a0 times a1;
run;
```

## 14. .

Describe precisely how you would use the least squares and logistic regression packages in SAS to estimate  $\theta_0, \theta_1, \theta_2$ , and  $\theta_3$  in the marginal structural model of Question 12 from the data described in Question 14. Again, write explicitly the SAS code, including the necessary SAS code to compute any weights you

might wish to use. (Here you are not to use your pseudo-population data, but rather the raw data. This question is to mimic Practicum I.)

SCIENTIST 1

Same answer as for question 14

SCIENTIST 2

The parameters cannot be estimated.

SCIENTIST 3

```
* estimate numerator of weights;
proc logistic data=name descending;
model a1=a0;
output out=data1 pred=num;
run;

* estimate denominator of weights;
proc logistic data=data1 descending;
model a1=a0 l1;
output out=data2 pred=den;
run;

* compute stabilized weights;
data final;
set data2;
if a1=1 then sw=num/den;
else if a1=0 then sw=(1-num)/(1-den);
run;

* estimation of MSM parameters;
proc reg data=final;
model y=a0 a1 a0a1;
weight sw;
run;
```

## 15. .

Is the MSM model of Question 12 correctly specified? How do you know? If your answer is yes and you are able to choose a SNM model in Question 3, explain how it is possible that both this MSM model is correct and the SNM model you chose in Question 3 is also correct.

SCIENTIST 1,2

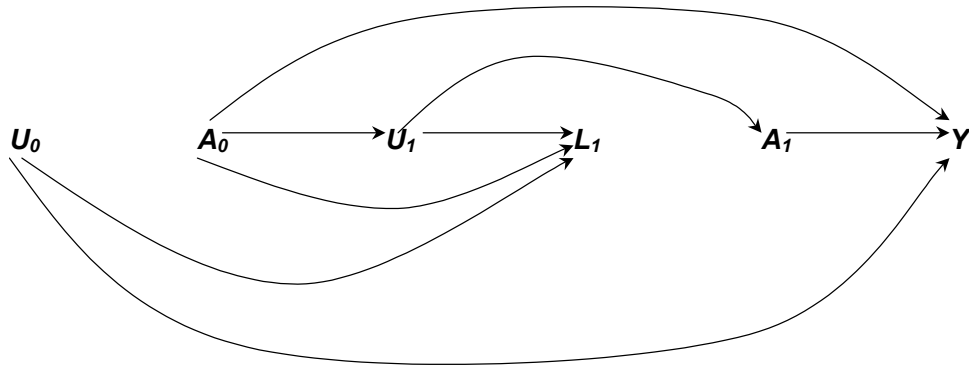
Yes, because it is a saturated model (i.e., it has enough parameters to reproduce exactly each of the four counterfactual means). No SNM chosen

SCIENTIST 3

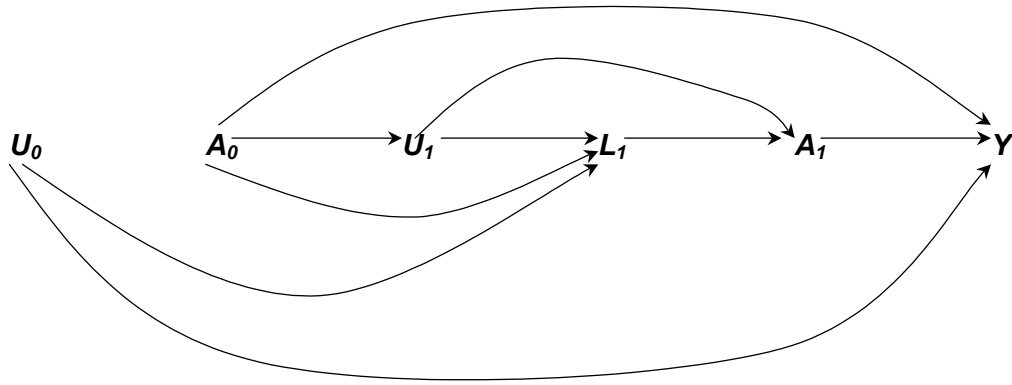
Yes, because it is a saturated model. The SNM 2.4 estimates how the the effect of  $A_1$  is modified by  $A_0$  and the time dependent covariate  $L_1$ , while the MSM estimates the total effect of the regimes  $g = (a_0, a_1)$  averaging over the differing effect of  $A_1$  within levels of  $A_1$  and  $L_1$ . Thus both models can be correct because the parameters represent different causal contrasts.

**Table 1: Data from an Observational Study**

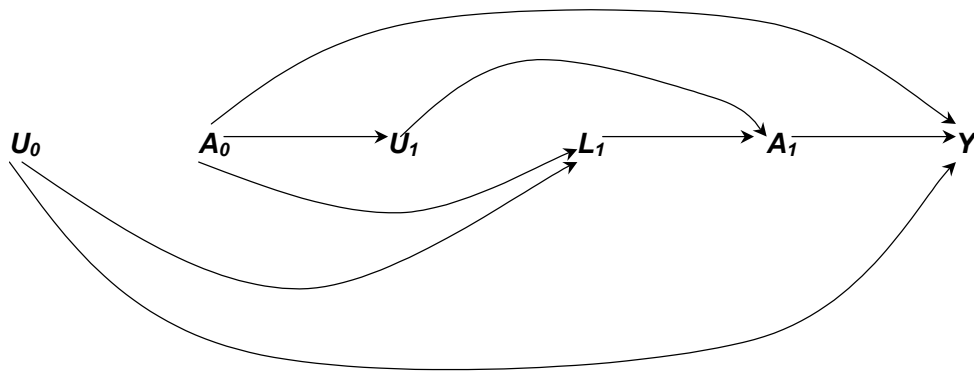
Row	$A_0$	$L_1$	$A_1$	No. of Subjects	$E[Y   A_0, L_1, A_1]$
1	0	1	0	2000	200
2	0	1	1	6000	220
3	0	0	0	6000	50
4	0	0	1	2000	70
5	1	1	0	3000	130
6	1	1	1	9000	110
7	1	0	0	3000	230
8	1	0	1	1000	250



DAG 1



DAG 2



DAG 3