MA592 Homework 4

1. Is $CD4_0$ a predictor of treatment A_0 in the first interval? Motivate your answer.

No, I don't think so. I'd say that the CD4_0 and A0 are independent. CD4_0 don't affect the treatment assignment in the first interval, the patients in A0 and A1 groups are both 500 no matter what is CD4_0 level.

2. Is $CD4_1$ a predictor of treatment A_1 in the second interval, for any given $CD4_0$ and A_0? Motivate your answer.

Yes. For given CD4_0 and A0, the treatment assignments are varied from A1 =0 and A1=1 groups in each CD4_1 level. In other words, the CD4_1 dose affect the A1.

3. Is $CD4_0$ a predictor of TD in the first interval for any of the treatments A_0 that started at baseline? Motivate your answer.

Yes. For A0=1, the TD in CD4_0 < = 200 level is higher than in CD4_0 > 200. For A0=0, the situation is the same, the TD in CD4_0 < = 200 level is also higher than in CD4_0 > 200. In conclusion, for any original treatment at baseline, CD4_0 affects the TD at the time point 1, so can be regarded as a predictor.

4. In this example, is the sample of patients initiating the 3-drug regimen representative of all patients included in this study? And what about the 4-drug regimen? Motivate your answer.

Yes. In the first time interval, there's no time-varying. The patients with 3-drug and 4-drug at baseline are similarly distributed in each CD4_0 level, patients with A0=1 and A0=0 are both 500 in each CD4_0 level. So, we can use patients initiating the 3-drug regimen representative of all patients included in this study.

5. In this example, is the sample of patients initiating the 3-drug regimen and continuing with the 3-drug regimen representative of all patients included in this study? Motivate your answer.

No. If the patients had TD in the time point 1, their target outcome are observed, so they would not participate in the second interval treatment assignment. These patients are very different with those who initiated and followed the 3-drug regimen throughout, since the latter are involved in the second treatment. These two populations are not comparable, so they cannot represent each other.

6. Suppose that your collaborator proposes to just use the proportion of patients with events among patients who are initiating the 3-drug regimen and to compare this with the proportion of patients with events among patients who are initiating the 4-drug regimen. What type of analysis would that be? Would this lead to a biased estimate of the probability of TD by time point 2 had everyone followed the 3-drug regimen throughout? Why/why not? Or, maybe more relevant, what question would this answer belong to?

1) type: intention-to-treat analysis.

2)bias: I think $\mathrm{E}(Y_1 + Y_2 | A0 = 1)$ would lead to biased estimator of $E(Y_2^{(1,1)})$. This estimation seems like ignore the CD4 level, since in A0=1 group, the proportion of CD4>200 always higher than in CD4<=200, the TD supposed to be less. So, the proportion of patients with events among patients who are initiating the 3-drug regimen underestimate the potential probability.

7. Suppose that we are interested in comparing the 3-drug regimen with the 4-drug regimen, had both of these treatment regimens been continued throughout after baseline. What could such analysis be called? Would the method proposed by your collaborator (see 6.) work to answer this question? Why/why not?

1)type: per-protocol analysis

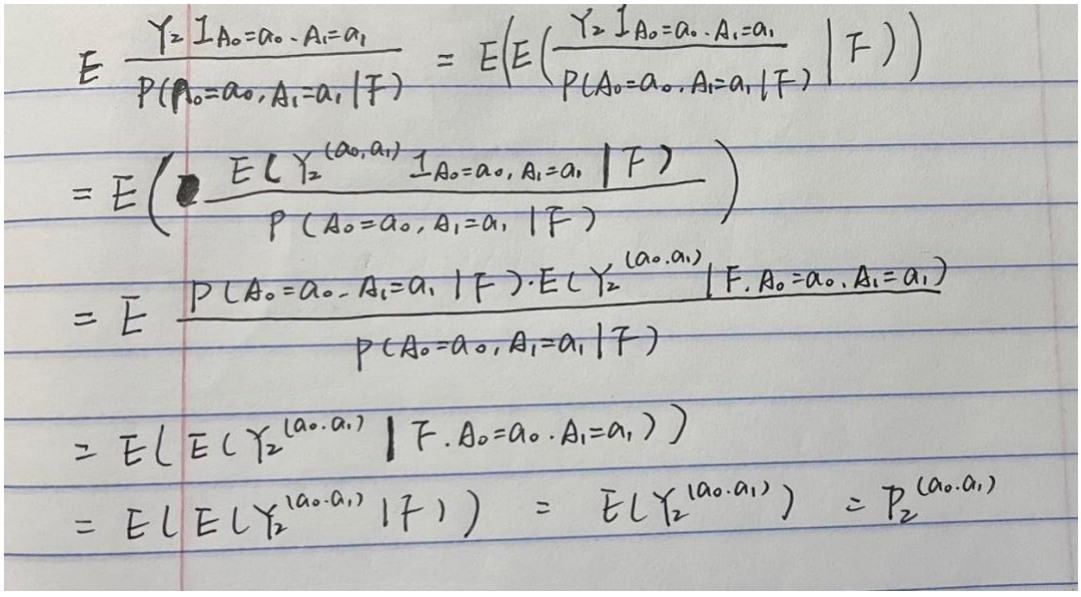
2) No. In the question 7 we want to get the results had all the patients followed the original treatment assignment at baseline. But in question 6, the method didn't care about whether the patients follow the original treatment throughout. It just used patients with A0=0 to estimate.

8. (9 points). Write $p_2^{(a_0,a_1)} = P\left(Y_{2i}^{(a_0,a_1)} = 1\right)$ for the probability of the outcome of Treatment Discontinuation by time point 2, under (a_0, a_1). Show that

$$E\left(\frac{1_{A_{0i}=a_0, A_{1i}=a_1}}{P(A_{0i}=a_0, A_{1i}=a_1|\mathbf{F}_i)}\left(Y_{2i} - p_2^{(a_0,a_1)}\right)\right) = 0.$$

Here,

$$1_{A_{0i}=a_0, A_{1i}=a_1} = \begin{cases} 1 & \text{if } A_{0i}=a_0, A_{1i}=a_1 \\ 0 & \text{if not.} \end{cases}$$

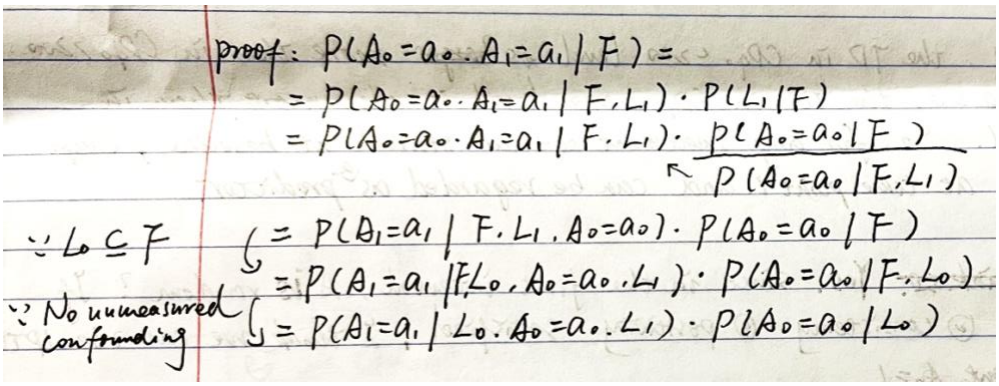


9. (8 points). The assumption of No Unmeasured Confounding in the time-varying setting is as follows:

$$A_{0i} \perp\!\!\!\perp \mathbf{F}_i | L_{0i}, \qquad A_{1i} \perp\!\!\!\perp \mathbf{F}_i | L_{0i}, A_{0i}, L_{1i}.$$

As for the time-fixed setting, this says that the treatment decisions may depend on a patient's past *observed* history, but not further on a patient's prognosis. Show that under No Unmeasured Confounding,

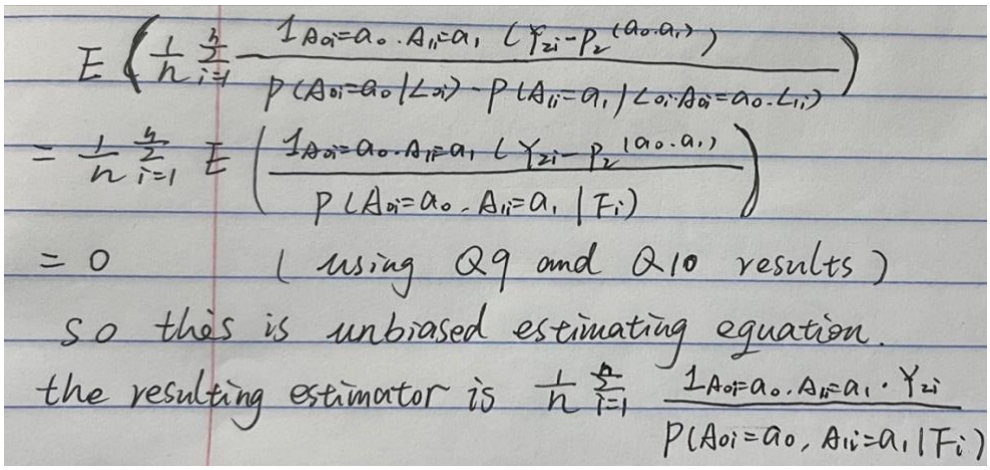
$$P(A_{0i} = a_0, A_{1i} = a_1 | \mathbf{F}_i) = P(A_{0i} = a_0 | L_{0i}) \cdot P(A_{1i} = a_1 | L_{0i}, A_{0i} = a_0, L_{1i}).$$



10. For a moment, assume that the treatment probabilities in 9. are known. Using 8. and 9., show that the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1_{A_{0i}=a_0, A_{1i}=a_1}}{P(A_{0i} = a_0 | L_{0i}) \cdot P(A_{1i} = a_1 | L_{0i}, A_{0i} = a_0, L_{1i})} \left(Y_{2i} - p_2^{(a_0,a_1)}\right) = 0$$

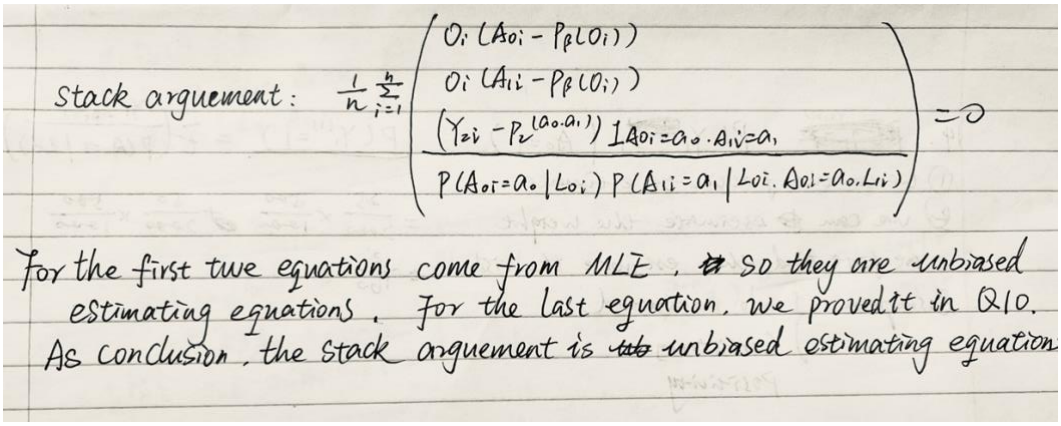
are unbiased estimating equations for $p_2^{(a_0,a_1)}$. What is the resulting estimator for $p_2^{(a_0,a_1)}$?



11. Using 10., how would you estimate $p_2^{(a_0,a_1)}$ if the treatment probabilities are unknown?

If the treatment probability $P(A_{0i} = a_0, A_{1i} = a_1|F_i)$ are unknown, we need to fit the logistics model to estimate it at first. And then put the $\hat{P}(A_{0i} = a_0, A_{1i} = a_1|F_i)$ into the equation in question 10 to calculate the $p_2^{(a_0,a_1)}$.

12. Using a stacking argument, show that your estimating equations from 11. are unbiased estimating equations for $p_2^{(a_0,a_1)}$.

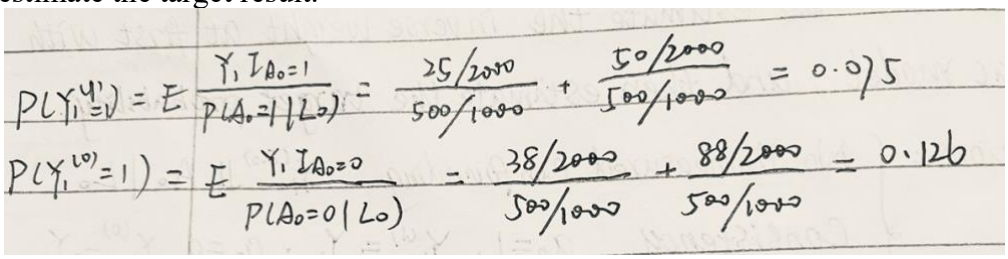


13. Which properties of your estimator for $p_2^{(a_0,a_1)}$ often follow from the fact that your estimator solves unbiased estimating equations? How would you estimate confidence intervals for your estimator for $p_2^{(a_0,a_1)}$ when the treatment probabilities are unknown and need to be estimated?

1) properties: consistency and asymptotic normality
2) CI: Use bootstrap. The bootstrap samples n patients with replacement say 2000 times, and in each dataset, we estimate the treatment effect using the respective method. The 95% confidence interval for the treatment effect is given by the 2.5% and 97.5% quantiles of the estimates based on the 2000 bootstrap samples.

14. How would you estimate the risk of Treatment Discontinuation (TD) by time point 1 in the population of the data described at the beginning of this assignment for the 3-drug regimen? What are your assumptions?

We can use fraction to estimate, or we can also fit logistics model to estimate the weight at first, and then estimate the target result.



Assumptions: { No Unmeasured confounding $Y_1^{(a_0)} \perp\!\!\!\perp A_0 \mid L_0$
Consistency $a_0=1,\ Y_1^{(1)} = Y_1$; $a_0=0,\ Y_1^{(0)} = Y_1$
Positivity $P(A_0=1 \mid L_0) \in (0,1)$

15. Pick two treatment regimes (a_0, a_1). Using the method you developed in 11., or the unbiased estimating equations from 10., estimate the probabilities of Treatment Discontinuation (TD) by time point 2, $p_2^{(a_0,a_1)}$, in the population of the data described at the beginning of this assignment. What are your assumptions? Hint: What is the weight of the $n = 75$ patients who had $CD4_0 > 200$, initiated the 3-drug regimen, and had TD=1 by time point 1? Their outcomes are observed, and their $A_{1i} = 0$, so their weight should not be 0 if you are considering a regime that starts with the 3-drug regimen.

I'm not sure about whether TD at time point 1 should be counted, so I wrote two methods:

Method1: Just consider TD in the time point 2:

Suppose $a_0=1, a_1=1$,

$$P(Y_2^{(1,1)}=1) = \frac{1}{N}\sum_{i=1}^{N} \frac{Y_2 \mathbb{1}_{A_0=1, A_1=1}}{P(A_0=1 \mid L_0) \cdot P(A_1=1 \mid L_0, A_0=1, L_1)}$$

$$= \frac{n_{L_0=0, L_1=0}/N}{P(A_0=1 \mid L_0=0) \cdot P(A_1=1 \mid L_0=0, A_0=1, L_1=0)} + \frac{n_{L_0=1, L_1=1}/N}{P(A_0=1 \mid L_0=1) \cdot P(A_1=1 \mid L_0=1, A_0=1, L_1=1)} + \frac{n_{L_0=1, L_1=0}/N}{P(A_0=1 \mid L_0=1)\, P(A_1=1 \mid L_0=1, A_0=1, L_1=0)}$$

total patients $N = 2000$

$$= \left(\frac{13}{\frac{1}{2} \times \frac{425}{475}} + \frac{16}{\frac{1}{2} \times \frac{175}{200}} + \frac{6}{\frac{1}{2} \times \frac{225}{250}} \right) / 2000$$

≈ 0.03948

Assumptions: { No Unmeasured Confounding
Consistency
Positivity

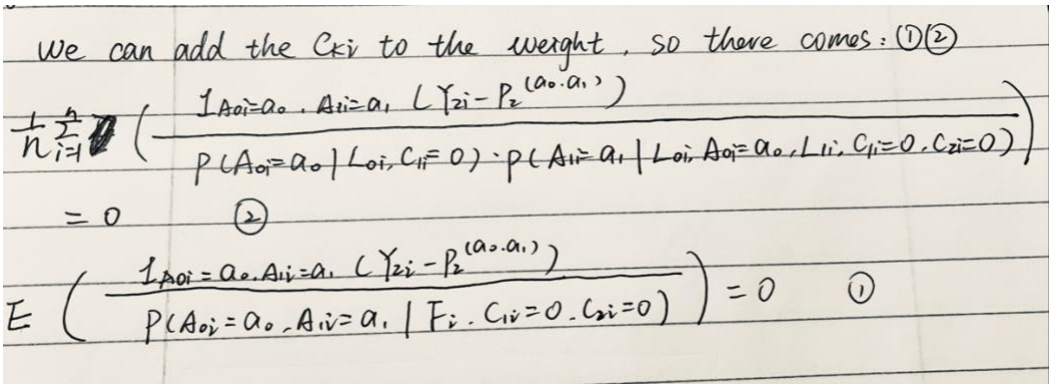
Method 2: Consider TD in the time point 1 and 2

If we consider $Y_1^{(1)}$, so we need to add the first two proportions:

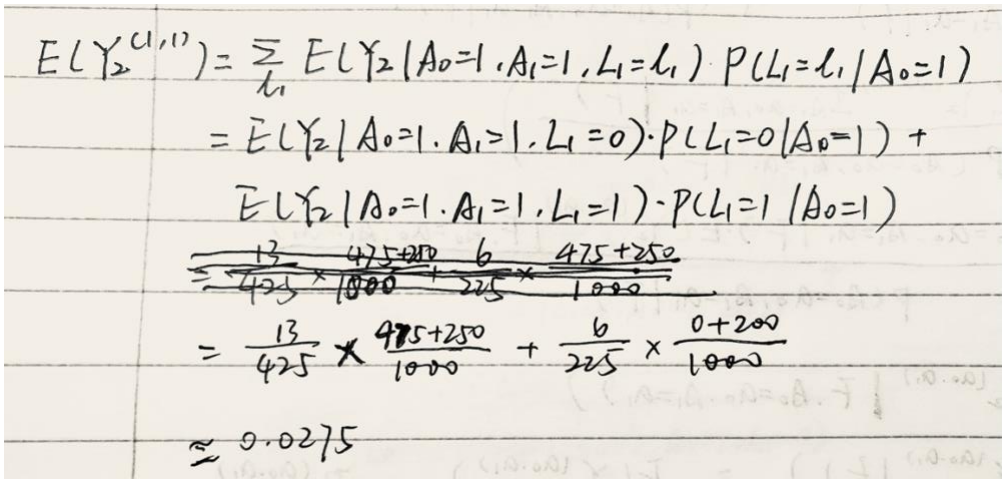
$$P(Y_2^{(1,1)}) = \left(\frac{25}{500/1000} + \frac{50}{500/1000} + \frac{13}{\frac{1}{2} \times \frac{425}{475}} + \frac{16}{\frac{1}{2} \times \frac{175}{200}} + \frac{6}{\frac{1}{2} \times \frac{225}{250}} \right) / 2000$$

≈ 0.11448

16. Suppose that additionally to the treatment variables A_{ki}, which may not be randomized, there is also censoring due to loss to follow-up. For $k = 1, 2$, let $C_{ki} = 1$ if patient i was censored before time point k, and $C_{ki} = 0$ if not. After an *observed* TD event, the censoring indicator will be 0. How would you go about incorporating censoring into the estimation process above? Just provide a sketch, you don't need to do everything all over again.



17. (bonus question). Can you also think of a conditioning argument (or maybe you need 2 conditioning arguments...) to estimate the probability of Treatment Discontinuation (TD) by time point 2 for any of the treatment regimens (a_0, a_1)?



I see this equation from the chapter 21 in the textbook *What If*, but I am not sure if it's correct.