Causal Inference with Structural Nested Models
Lab, Catalyst, 11/22/2019
Coarse Structural Nested Mean Models

Judith J. Lok

November 21, 2019

We consider an observational HIV study where the outcome of interest is the CD4 count. The data have been simulated based on the AIEDRP data, see e.g. Hecht et al. (2006), Lok and DeGruttola (2012) and Yang et al. (2016).

For this simulation, we have restricted the data to 6-18 months after estimated date of HIV infection. More about the simulated data has been presented in the course slides. The simulated dataset is called simAIEDRP.

1. Postulate a model for how the probability of treatment initiation, given that treatment was not initiated before, depends on injection drug use, month, and current CD4 count. You may want to postulate a separate model for the first time point, where prior data are not available, and for the later time points, where time-dependent data is available.

2. Postulate one overall model for the probability of treatment initiation, for all time points, including the baseline and later time points, without changing the assumptions of the model in 1. Hint: you may want to use interactions with $1_{k>6}$.

3. Using the simulated data, fit the model you postulated in 2. If you would like to check your answer to 2., you may also want to fit the models you postulated in 1. and compare the results.

4. For a given treatment, do later CD4 counts depend on the variables which are predictors of treatment initiation such as the current CD4 count? How do you find out?

5. Based on these data, do you think there is confounding by indication?

6. In this assignment, we focus on estimating how the mean of Y_{30} depends on treatment initiation. Using the simulated data and your model from 3., test H_0: Y_{30} doesn't depend on treatment initiation.

7. Carry out a naive analysis, predicting the mean CD4 count in month 30 based on injection drug use, $CD4_6$, and the month of treatment initiation:

$$Y_{30} = \beta_0 + \beta_1 injdrug + \beta_2 CD4_6 + \psi Tr(30),$$

where $Tr(30)$ is the duration of treatment between month 6 and month 30 (this is 0 for patients who never initiated treatment). Is this a valid analysis?

8. Consider the 1-dimensional model for the effect of treatment on Y_{30},

$$\gamma^{30}_{m,\psi}\left(\bar{l}_m\right) = \psi_1(30 - m), \tag{1}$$

$(30 - m)$: duration of treatment from month m to month 30. That is, the effect of treatment on the outcome at month 30 is linear in its duration. Fit this model in the simulated data. Hint 1: the following theorem might help:

Theorem *(Unbiased Estimating Equations).* For model (1), the theorem from slide 97 becomes:

$$E\sum_{m=0}^{29} \vec{q}_m\left(\overline{L}_m\right) H\left(30\right) 1_{\overline{A}_{m-1}=\bar{0}} 1_{visit}(m)\left(A_m - p\left(m\right)\right) = 0,$$

with $1_{visit}(m)$ the indicator of whether a visit took place at month m. Thus if $p_\theta\left(m\right)$ and γ_ψ are correctly specified (parametric) models for $p\left(m\right)$ and γ, then

$$P_n\sum_{m=0}^{29} \vec{q}_m\left(\overline{L}_m\right) H_\psi\left(30\right) 1_{\overline{A}_{m-1}=\bar{0}} 1_{visit}(m)\left(A_m - p_\theta\left(m\right)\right) = 0,$$

with P_n the empirical measure $P_n X = 1/n\sum_{i=1}^n X_i$, are unbiased estimation equations for both ψ^* and the (nuisance) parameter θ^*.

Hint 2: for this model, if Tr_{30} is the duration of treatment until month 30 (0 if the patient was not treated), $H_\psi\left(30\right) = Y_{30} - \psi_1 Tr_{30}$. You could approach this in 2 ways:

(a) Use the linearity (in ψ) of $\gamma^{30}_{m,\psi}$ (and hence H_ψ) to solve the estimating equations for ψ directly after calculating the appropriate averages. Important hint: it might help to choose $q_m(L_m) = CD4_m$. AND/OR:

(b) Create a grid of potential ψ's, such as $[-10, 40]$. Calculate $H_\psi(30)$ for the ψ's on the grid. Add each of the $H_\psi(30)$ to the model for treatment initiation, separately. $\hat{\psi}$ is the ψ for which $\hat{\alpha}(\psi) = 0$. You may have to add a few more grid points where $\hat{\alpha}(\psi)$ is close to 0, or do interpolation to obtain a more precise estimate. Important hint: instead of adding $H_\psi(30)$ directly, it may be better to add $CD4_m \cdot H_\psi(30)$ to the model for treatment initiation.

9. Estimate the 2-dimensional model

$$\gamma^{30}_{m,\psi}\left(\bar{l}_m\right) = \left(\psi_1 + \psi_2 m\right)(30 - m),$$

$(30 - m)$: duration of treatment from month m to month 30. Effect of treatment linear in its duration. Coefficient depends on month of treatment initiation. In this case,

$$\begin{aligned} H_\psi(30) &= Y_{30} - \psi_1 \cdot Tr_{30} - \psi_2 \cdot \text{monthfirsttrt} \cdot Tr_{30} \\ &= Y_{30} - \left(\begin{array}{cc} Tr_{30} & \text{monthfirsttrt} \cdot Tr_{30} \end{array} \right) \left(\begin{array}{c} \psi_1 \\ \psi_2 \end{array} \right). \end{aligned}$$

You could approach this in 2 ways:

(a) Use the linearity (in ψ) of $\gamma^{30}_{m,\psi}$ (and hence H_ψ) to solve the estimating equations for ψ directly after calculating the appropriate averages. Try e.g. $\vec{q}_m\left(\overline{L}_m\right) = (CD4_m, m)^\top$. In R, calculating matrix inverses is easy. If you use SAS, you can either use PROC IML or calculate the required matrix inverse directly; in case you forgot:

$$\left(\begin{array}{cc} a & b \\ c & d \end{array} \right)^{-1} = \left(\begin{array}{cc} d & -b \\ -c & a \end{array} \right) / (ad - bc).$$

AND/OR:

(b) Create a grid of potential ψ's, such as $[-10, 40]$ for ψ_1 and $[-1, 1]$ for ψ_2. Calculate $H_\psi(30)$ for the ψ's on the grid. Add each of the H_ψ to the model for treatment initiation, separately. $\hat{\psi}$ is the ψ for which $\hat{\alpha}(\psi) = 0$. You may have to add a few more grid points where $\hat{\alpha}(\psi)$ is close to 0. Important hint: you will need to add a 2-dimensional version of $H_\psi(30)$ to the model for treatment initiation! E.g., $(CD4_m \cdot H_\psi(30), m \cdot H_\psi(30))$.

10. Compare your coarse SNMM analyses with the naive analysis in Question 7.

11. Extra question: Using similar techniques, you could also estimate the model

$$\gamma^k_{m,\psi}\left(\bar{l}_m\right) = (\psi_1 + \psi_2 m)(k - m), \tag{2}$$

for $k \geq m$ and $k \leq m + 12$.

References

Hecht, F. M., L. Wang, A. Collier, S. Little, M. Markowitz, J. Margolick, J. M. Kilby, E. Daar, B. Conway, S. Holte, and AIEDRP Network (2006). A multicenter observational study of the potential benefits of initiating combination antiretroviral therapy during acute HIV infection. *Journal of Infectious Disease 194*, 725–733.

Lok, J. J. and V. DeGruttola (2012). Impact of time to start treatment following infection with application to initiating HAART in HIV-positive patients. *Biometrics 68*, 745—-754.

Yang, S., S. Little, D. Smith, V. DeGruttola, and J. J. Lok (2016). Impact of each month delay in initiation on the effect of 1 year of ART on CD4 count. In *Conference on Retroviruses and Opportunistic Infections (CROI)*. Poster abstract 16-647.