Homework Assignment 4

Statistical Methods for Analysis with Missing Data, Winter 2019

Instructor: Mauricio Sadinle, Department of Biostatistics, U. of Washington – Seattle

Submit your solutions via Canvas. Due by 12:00pm (noon) on March 20, 2019.

From this assignment you can get a maximum of 20 points. The assignment contains a list of problems, each worth a different number of points. You may choose any combination of problems that you like. I recommend that you solve a combination of problems that is worth more than 20 points as a way of gaining insurance against errors in some of your problem solutions. If you are submitting solutions to theoretical problems, feel free to hand-write them and submit a scanned copy.

For problems 1 – 3, consider a scenario where covariates X are fully observed, and outcome Y might be missing, with R representing its response indicator. Say that using i.i.d. data $\{(X_i, Y_{i(R_i)}, R_i)\}_{i=1}^n$ you estimate the propensity score $p(R = 1 \mid x)$ as $\pi(x; \hat{\psi})$, where $\hat{\psi}$ is a vector of estimated parameters, e.g. using a logistic regression of R on X. We will be assuming that $R \perp\!\!\!\perp Y \mid X$. Our goal is to estimate $\mu = E(Y)$. Remember from problem 11 of HW1 that the expected value of the complete case estimator of μ

$$\hat{\mu}^{cc} = \frac{\sum_{i=1}^n R_i Y_i}{\sum_{i=1}^n R_i}$$

is $E(Y \mid R = 1)$, conditioning on at least one observed response. This means that $\hat{\mu}^{cc}$ is generally biased for μ, and it motivates the usage of the inverse-probability weighted estimator

$$\hat{\mu}^{ipw} = \frac{1}{n} \sum_{i=1}^n \frac{R_i}{\pi(X_i; \hat{\psi})} Y_i.$$

1. (5 points, skip if you are not familiar with consistency and convergence in probability) Show that $\hat{\mu}^{ipw}$ is consistent for μ if $\pi(x; \psi)$ is correctly specified, that is, if there exists ψ^* such that $\pi(x; \psi^*) = p(R = 1 \mid x)$, and if the estimator $\hat{\psi}$ consistently estimates ψ^*, i.e., if $\hat{\psi} \xrightarrow{p} \psi^*$.

2. (5 points, skip if you are not familiar with consistency and convergence in probability) Now, additionally consider a conditional mean model $m(x;\xi)$ for $E(Y \mid x)$. The *augmented IPW* (AIPW) estimator of the population mean is

$$\hat{\mu}^{aipw} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\pi(X_i;\hat{\psi})} - \frac{1}{n}\sum_{i=1}^{n} \frac{(R_i - \pi(X_i;\hat{\psi}))}{\pi(X_i;\hat{\psi})} m(X_i;\hat{\xi}).$$

Show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R - \pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right).$$

(Note that for this problem we are not assuming that the propensity score and conditional mean models are correctly specified).

3. (5 points) In the previous problem, we saw that as the sample size increases the estimator $\hat{\mu}^{aipw}$ converges to the quantity

$$E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R - \pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right).$$

Show that this can be rewriten as

$$\mu + E_X\left[E_R\left(\frac{(R - \pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

and explain what happens if either the propensity score or the conditional mean models are correctly specified, that is, if ψ^* is such that $\pi(x;\psi^*) = p(R = 1 \mid x)$ or if ξ^* is such that $m(x;\xi^*) = E(Y \mid x)$.

For problems 4 – 9 consider a vector of measurements $Z = (Y_1, \ldots, Y_T)$ taken over T time periods, and their response indicators $(R_1, \ldots, R_T)$. We will work under dropout, where $R_j = 0$ implies that $R_{j'} = 0$ for all $j' > j$. Let D be the *dropout* indicator, where $D = j + 1$ indicates that the individual is last seen at time j.

4. (1 point) Write the formula for D in terms of the response indicators $(R_1, \ldots, R_T)$.

5. (1 point) Denote the dropout hazard function as $\lambda_j(Z) = p(D = j \mid D \geq j, Z)$. Find an equivalent expression for $\lambda_j(Z)$ in terms of the response indicators $(R_1, \ldots, R_T)$.

6. (2 points) Show that $p(D = j + 1 \mid Z) = \lambda_{j+1}(Z) \prod_{\ell=1}^{j}[1 - \lambda_\ell(Z)]$

7. (2 points) Show that $p(R_j = 1 \mid Z) = \prod_{\ell=1}^{j}[1 - \lambda_\ell(Z)]$

8. (3 points) Show that the MAR assumption in this case is equivalent to assuming

$$\lambda_j(Z) = p(D = j \mid D \geq j, Z) \overset{MAR}{=} p(D = j \mid D \geq j, H_{j-1}) = \lambda_j(H_{j-1}),$$

where $H_j = (Y_1, \ldots, Y_j)$.

9. (1 point) Explain how each $\lambda_j(H_{j-1})$ can be directly estimated from the observed data.

The following is a computational problem that builds on R session 4.

10. (10 points) One of the appeals of (weighted) generalized estimating equations is that they lead to consistent estimators even when the working correlation matrix is misspecified (provided correct specification of the propensity score model and the marginal response models for each time period). The goal of this problem is to explore this property using the R package `wgeesel`. The idea is to run a simulation study where you generate multiple datasets using the function `data_sim` under an exchangeable correlation structure. For each of these datasets, use the function `wgee` to fit WGEEs with correctly specified propensity score and marginal response models, but changing the working correlation structure to exchangeable, AR1 and unstructured. Repeat your simulation study under three different set-ups for the number of individuals: 100, 500, and 2000. For other details on how to set-up your simulation study, follow the set-up of the simulation of Section 4 in the `wgeesel` tutorial. Report your results in a table similar to Table 2 of the `wgeesel` tutorial, and also submit your code.

The points obtained from problems 11 – 14 will be added to the score of any of your previous homework solutions in which you obtained less than 20 points. They can also be used towards the 20 points from this homework.

11. (1 point) Say we have a longitudinal study with dropout and $T = 3$. Write down the neighboring-case identifying assumption for $\ell \geq d$, $d = 1, 2, 3$.

12. (1 point) Say we have a longitudinal study with dropout and $T = 3$. Write down the available-case identifying assumption for $\ell \geq d$, $d = 1, 2, 3$.

13. (1 point) Under monotone nonresponse (e.g., dropout), show that the available-case identifying assumption is equivalent to MAR.

14. (1 point) Show that the full-data distributions obtained under the complete-case, neighboring-case, and available-case identifying assumptions are observationally equivalent (under dropout) and nonparametric identified.