Homework Assignment 1

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 Jan 30, 2019.

From this assignment you can get a maximum of 20 points. The assignment contains a long 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.

In problems 1–3, Y is a random variable and R its response indicator. We denote the conditional densities as $p(y \mid R = r), r = 0, 1$, the marginal density of Y as $p(y)$, and likewise for other marginal and conditional densities.

1. (1 point) Show that assuming that $p(y \mid R = 0) = p(y \mid R = 1)$ is equivalent to assuming $p(y \mid R = 1) = p(y)$.

2. (1 point) Show that assuming that $p(y \mid R = 0) = p(y \mid R = 1)$ is equivalent to assuming $p(R = r \mid y) = p(R = r)$ for all y, $r = 0, 1$.

3. (3 points) If $p(R = 1 \mid y)$ is an increasing function of y, show that

$$E(Y \mid R = 1) > E(Y).$$

For problems 4–9, $Z = (Z_1, \ldots, Z_K)$ and $R = (R_1, \ldots, R_K)$.

4. (1 point) Say $K = 3$. Write down $Z_{(r)}$ and $Z_{(\bar{r})}$ for all possible values of $r \in \{0, 1\}^3$.

5. (1 point) Explain what is the difference between $(Z_{(R)}, R)$ and $(Z_{(r)}, R = r)$ for a fixed value r.

6. (1 point) Say $K = 2$, $Z_1 \in \{1, 2\}$, $Z_2 \in \{A, B\}$, $R \in \{0, 1\}^2$. Write down all the elements of the sample space of $(Z_{(R)}, R)$.

7. (1 point) Say $K = 2$. Describe the sample space of $(Z_{(R)}, R)$ if Z_1 and Z_2 take values in the real numbers.

8. (1 point) Data are said to be *missing at random* (MAR) if $p(R = r \mid z) = p(R = r \mid z_{(r)})$. Say $K = 3$. Write down the MAR assumption for each individual $r \in \{0, 1\}^3$.

9. (1 point) Data are said to be *missing completely at random* (MCAR) if $p(R = r \mid z) = p(R = r)$. Show that MAR and MCAR are equivalent when $K = 1$.

For problems 10–16, say you take a random sample $\{(Y_i, R_i)\}_{i=1}^n \overset{i.i.d.}{\sim} F$, where F is the joint distribution of some generic random variables Y and R. You don't get to see Y_i when $R_i = 0$. For all of your proofs, condition on at least one Y_i being observed. Y is numeric.

10. (1 point) Explain why $p(R = 1 \mid y)$ cannot be estimated from the observed data.

11. (2 points) Consider the complete-case estimator of the mean:

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

Show that $E(\hat{\mu}^{cc}) = E(Y \mid R = 1)$ for all sample sizes.

12. (3 points) Compute the (theoretical) variance of $\hat{\mu}^{cc}$.

13. (5 points) A possible estimator of the variance of Y is

$$\hat{V}^{cc}(Y) = \frac{\sum_{i=1}^n (Y_i - \hat{\mu}^{cc})^2 R_i}{\sum_{i=1}^n R_i}.$$

Compute its expected value.

14. (5 points) A possible estimator of the variance of $\hat{\mu}^{cc}$ is

$$\hat{V}^{cc}(\hat{\mu}^{cc}) = \frac{\sum_{i=1}^n (Y_i - \hat{\mu}^{cc})^2 R_i}{(\sum_{i=1}^n R_i)^2}.$$

Compute its expected value.

15. (1 point) Estimating the mean after mean imputation corresponds to using the estimator

$$\hat{\mu}^{mimp} = \frac{1}{n}\sum_{i=1}^{n}[Y_i R_i + \hat{\mu}^{cc}(1 - R_i)].$$

Show that $\hat{\mu}^{mimp} = \hat{\mu}^{cc}$.

16. (1 point) Comment on the implications of single imputation for the construction of confidence intervals.

The following are computational problems that build on R session 1.

17. (10 points) Design and run a simulation study with the goal of exploring the performance of hot-deck imputation in terms of estimation of regression coefficients and means. Compare your results with mean imputation and complete case analysis. Submit your R code and pdf report with your results. If you plan to solve this problem, consult with me for guidance.

18. (20 points) Design and run a simulation study with the goal of exploring the performance of the bootstrap plus hot-deck imputation in terms of estimation of regression coefficients and means. Compare your results with mean imputation and complete case analysis. Submit your R code and pdf report with your results. If you plan to solve this problem, consult with me for guidance.