Assignment 1

BIOST/EPI 531

Due Feb 1, 2017

1. This is a hypothetical example to improve your understanding of the missing data mechanisms. Suppose that you want to investigate the quality of life of a group of 200 cancer patients. Based on self report, the patients classified themselves into "good", "moderate" and "bad" quality of life. The histologic grade of cancer are classified into "high" and "low" grades. Table 1 provides the data that you would observe when there is no missing data.

   (a) Use the data in Table 1, estimate the marginal distribution of quality of life (i.e. the percentage in each category). Estimate the conditional distribution of quality of life given "high" grade histology, and the conditional distribution of quality of life given "low" grade histology. Are they different? Does this make sense?

   Now suppose that you know the histologic grade of cancer for all subjects, but some quality-of-life ratings were missing. Tables 2-4 provide three different data that you may observe.

   (b) What does it mean to say that quality of life is missing completely at random, missing at random and missing not at random? Explain in terms of the variables being collected.

   (c) For each of the three observed data (Table 2, 3, 4), describe the missing data mechanisms. Use complete-case analysis to estimate the marginal distribution of quality of life (i.e. the percentage in each category), the conditional distribution

of quality of life given "high" grade histology, and the conditional distribution of quality of life given "low" grade histology. Which estimates are biased and which are not?

(d) For each of the three observed data (Table 2, 3, 4), use inverse probability weighting to estimate the marginal distribution of quality of life. Are your estimates unbiased?

2. Download data "mental_baseline.csv". It contains the baseline data of a mental health trial. The data set contains 3 variables: hospday: days in hospital for psychiatric reasons, cprs: a measure of psychopathology, sat: a measure of patient's satisfaction with mental heath services.

   (a) Describe the missing data pattern.

   (b) Is the missingness mechanism MCAR? Explain your conclusions based on a logistic regression model of the missingness events.

   (c) Estimate the mean of "sat" using complete-case analysis and weighted complete-case analysis with inverse probability weighting. Obtain 95% confidence intervals for your estimates. (For IPW estimation, you need to obtain 95% CI through bootstrap)

   (d) Explain the reasons that the two estimates are similar. (Hint: check if the variable "sat" is associated with "hospday")

3. Download data "bc.csv". It contains data from an observational study conducted by the German Breast Cancer Study Group. The data set contains 4 variables. tmass: tumor size in mm at the study endpoint; treatment: 1=mastectomy, 0=breast conservation; tgr: tumor size before treatment (1: greater than 10mm, 0: otherwise); age: age in years at treatment.

   (a) Use a logistic regression model to estimate the propensity score of receiving mastectomy.

   (b) Estimate the average treatment effect, that is, the difference of average tumor size between mastectomy and breast conservation, using

i. Empirical mean difference (without any weighting or adjustment)

ii. Inverse probability weighting

iii. Nonparametric calibration estimator

(c) Compare your findings from (b).

4. Briefly describe what you would investigate in your term project. We have two possible options for a final project.

   (a) If you have a data set in hand with missing data, you can write an analysis proposal describing the data and how you could analyze the data. Explain the pros and cons of your proposal.

   (b) You can also focus on a published paper with missing data analysis, describe what the authors have done and the underlying assumptions for the analysis. Critique the method.

   If you proceed with option (a), please provide a brief description of the data set you have and its missing data pattern. Otherwise, please identify a paper in biostatistics and epidemiology literature that contains analysis of missing data. Please give the citation of the paper.

Table 1: Complete data

	Quality of life			
	good	moderate	poor	Total
Low grade	80	40	12	132
High grade	8	20	40	68
Total	88	60	52	200

Table 2: Missing data case 1

	Quality of life				
	good	moderate	poor	missing	Total
Low grade	60	20	3	49	132
High grade	6	10	10	42	68
Total	66	30	13	91	200

Table 3: Missing data case 2

	Quality of life				
	good	moderate	poor	missing	Total
Low grade	60	30	9	33	132
High grade	6	15	30	17	68
Total	66	45	39	50	200

Table 4: Missing data case 3

	Quality of life				
	good	moderate	poor	missing	Total
Low grade	60	30	9	33	132
High grade	4	10	20	34	68
Total	64	40	29	67	200