UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Assignment 1

- Deadline for submission: 4pm, February 16, 2023.
- Location for submission: Gradescope over Learn. **Important**: When uploading your report to Gradescope please tag separately each subquestion (e.g. 1a), 1b), 1c), etc).
- This assignment is worth 20% of your final grade for the course.
- Assignments should be typed (LaTeX, word, etc.).
- Answers to questions should be in full sentences and should provide all necessary details.
- Any output (e.g., graphs, tables) from `R` that you use to answer questions must be included with the assignment. Also, please include in the pdf of your submission the code for questions 3 and 4. Pdfs generated via `R` markdown are welcome.
- The assignment is out of 100 marks.

1. Let us suppose that on a (hypothetical) survey there is a question about alcohol consumption and that the variable ALQ records the respondent's answer to the following specific question: "In the past year, have you had at least 12 drinks of any type of alcoholic beverage?". The possible answers are 'Yes' or 'No'. Not all participants respond to this question, that is, the ALQ variable has some missing values. Further, and again hypothetically, suppose that we only have additional data on the gender of the participants in the survey (which is fully observed). For each of the following situations, choose, justifying, the correct answer.

   (a) Suppose that ALQ is MCAR. The probability of ALQ being missing for those with ALQ=Yes is 0.3. What is the probability of ALQ being missing for those with ALQ=No?

   (i) 0.03
   (ii) 0.3
   (iii) 0.33

   (**5 marks**)

   (b) ALQ being MAR given gender means:

   (i) The probability of ALQ being missing depends on the Yes/No value of ALQ even after adjusting for gender.

(ii) The probability of ALQ being missing is independent of the Yes/No value of ALQ after adjusting for gender.

(iii) The probability of ALQ being missing is independent of the Yes/No value of ALQ and gender.

(**5 marks**)

(c) Suppose again that ALQ is MAR given gender, and that the probability of ALQ being missing for men is 0.1. What is the probability of ALQ being missing for women?

(i) 0.1

(ii) 0.9

(iii) It is impossible to conclude from the information given.

(**5 marks**)

2. Suppose that a dataset consists of 100 subjects and 10 variables. Each variable contains 10% of missing values. What is the largest possible subsample under a complete case analysis? What is the smallest? Justify. (**10 marks**)

3. Consider a two variable (Y_1, Y_2) problem, with each variable defined as follows:

$$Y_1 = 1 + Z_1,$$
$$Y_2 = 5 + 2 \times Z_1 + Z_2,$$

where Y_1 is fully observed but Y_2 is subject to missingness. Further consider that Y_2 is missing if $a \times (Y_1 - 1) + b \times (Y_2 - 5) + Z_3 < 0$, where Z_1, Z_2, and Z_3 follow independent standard normal (that is, mean 0 and variance 1) distributions. **Important**: Please use `set.seed(1)` in R when simulating the data, for reproducibility reasons.

(a) Start by simulating a (complete) dataset of size 500 on (Y_1, Y_2). Then, and considering $a = 2$ and $b = 0$, simulate the corresponding observed dataset (by imposing missingness on Y_2 as instructed above). Is this mechanism MCAR, MAR, or MNAR? Display the marginal distribution of Y_2 for the complete (as originally simulated) and observed (after imposing missingness) data. Comment. (**10 marks**)

(b) For the observed dataset simulated in (a), impute the missing values using stochastic regression imputation. Display the marginal distribution of Y_2 for the complete (as originally simulated) and *completed* (after imputation) data. Comment. (**5 marks**)

(c) Using the complete dataset simulated in (a), now impose missingness on Y_2 by considering $a = 0$ and $b = 2$. Is this mechanism MCAR, MAR, or MNAR? Display the marginal distribution of Y_2 for the complete (as originally simulated) and observed (after imposing missingness) data. Comment. (**10 marks**)

(d) The same as in (b) but for the observed data generated in (c). (**5 marks**)

4. It is sometimes necessary to lower a patient's blood pressure during surgery, using a hypotensive drug. Such drugs are administrated continuously during the relevant phase of the operation; because the duration of this phase varies, so does the total amount of drug administered. Patients also vary in the extent to which the drugs succeed in lowering blood pressure. The sooner the blood pressure rises again to normal after the drug is discontinued, the better. The dataset `databp.Rdata` available on Learn, a partial missing value version of the data presented by Robertson and Armitage (1959), relate to a particular hypotensive drug and give the time in minutes before the patient's systolic blood pressure returned to 1000mm of mercury (the recovery time), the logarithm (base 10) of the dose of drug in milligrams (you can use this variable as is, no need to transform it to the original scale), and the average systolic blood pressure achieved while the drug was being administered.

   (a) Carry out a complete case analysis to find the mean value of the recovery time (and associated standard error) and to find also the (Pearson) correlations between the recovery time and the dose and between the recovery time and blood pressure. (**5 marks**)

   (b) The same as in (a) but using mean imputation. (**5 marks**)

   (c) The same as in (a) but using mean regression imputation. (**5 marks**)

   (d) The same as in (a) but using stochastic regression imputation. Do you need any extra care when conducting stochastic regression imputation in this example? (**5 marks**)

   (e) You will now conduct the same analysis but applying another technique called predictive mean matching (Little, 1988), which is a special type of hot deck imputation. In the simplest form of this method (and the one you will use here), a regression model is used to predict the variables with missing values from the other (complete) variables. For each subject with a missing value, the donor is chosen to be the subject with a predicted value of her or his own that is closest (to be measured by the squared difference) to the prediction for the subject with the missing value. (**20 marks**)

   (f) What is an advantage of predictive mean matching over stochastic regression imputation? Based on your analysis, can you foresee any potential problem of predictive mean matching? (**5 marks**)

References

Little, R. J. (1988). Missing data adjustments in large surveys. *Journal of Business and Economic Statistics* **6**, 287–296.