UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Assignment

- To be uploaded to Learn by 23:59, Sunday November 10, 2019.
- This assignment is worth 5% of your final grade for the course.
- Assignments should be typed (LaTeX, word etc.).
- Answers to questions should be in full sentences.
- Any output (e.g., graphs, tables) from `R` that you use to answer questions must be included with the assignment. Also, please append your `R` code at the end of the assignment or make it available in a public repository.
- The assignment is out of 100 marks.

1. The definition of MAR can depend on how the complete data are defined. Suppose that $Y = (Y_1, \ldots, Y_n)$ is a normally distributed random sample with mean μ_Y and variance σ_Y^2, and $Z = (Z_1, \ldots, Z_n)$ are completely unobserved variables, which are independent of Y. Suppose that some values Y_i are missing and

$$\Pr(R_i = 1 \mid Y, Z) = \frac{\exp(Z_i)}{1 + \exp(Z_i)}.$$

   where R_i, $i = 1, \ldots, n$, is the missing value indicator, taking the value 1 if the corresponding Y_i value is observed and 0 otherwise. Argue that if the complete data are defined as Y then the missing data mechanism satisfies MCAR, but if the complete data are defined as (Y, Z), then the missing observations are not MAR. Which is the more sensible definition in this case? (**30 marks**)

2. 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. (**10 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? (**10 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. (**30 marks**)

(f) What is an advantage of predictive mean matching over stochastic regression imputation? Can you foresee any potential problem of predictive mean matching? (**10 marks**)

References

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