UNIVERSITY OF EDINBURGH

SCHOOL OF MATHEMATICS

INCOMPLETE DATA ANALYSIS

Assignment

- To be uploaded to Learn by 23:59, Sunday November 18, 2018.

- 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.

- The assignment is out of 100 marks.

- You are expected to work independently and not to discuss the assignment with others.

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, and the average systolic blood pressure achieved while the drug was being administered.

(a) (**10 marks**) Carry out a complete case analysis to find the mean value of the recovery time (and associated standard error) and to find also the correlations between the recovery time and the dose and between the recovery time and blood pressure.

(b) (**15 marks**) The same as in (a) but using mean imputation.

(c) (**15 marks**) The same as in (a) but using mean regression imputation.

(d) (**15 marks**) The same as in (a) but using stochastic regression imputation. Do you need any extra care when conducting stochastic regression imputation in this example?

(e) (**30 marks**) 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.

(f) (**15 marks**) What is an obvious advantage of predictive mean matching over stochastic regression imputation?