UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Two practice exercises

1. Consider a two variable (Y_1, Y_2) data matrix with only Y_2 subject to missingness. Conducting a complete case analysis, one has obtained that the estimated mean of Y_2 is 35. For this same dataset, if we were to conduct a mean imputation approach instead of a complete case analysis, what would be the estimated mean of Y_2 after imputation? Justify.

2. The following table shows a small artificial dataset with 4 missing values:

Subject number	Sex	Age(years)	Systolic blood pressure (mmHg)
1	Male	50	163.5
2	Male	41	126.4
3	Male	52	150.7
4	Male	58	190.4
5	Male	56	172.2
6	Male	45	NA
7	Male	42	136.3
8	Male	48	146.8
9	Male	57	162.5
10	Male	56	161.0
11	Male	55	148.7
12	Male	58	163.6
13	Female	57	NA
14	Female	44	140.6
15	Female	56	NA
16	Female	45	118.7
17	Female	48	NA
18	Female	50	104.6
19	Female	59	131.5
20	Female	55	126.9

The file `datasbp.Rdata` is available on Learn. Save the data to your preferred directory and then read it in `R` by typing `load("datasbp.Rdata")` (once you are in the correct directory in `R`).

(a) Carry out a complete case analysis to find the mean value of systolic blood pressure overall, and by gender. Also compute the associated standard error of the mean.

(b) Impute the missing values of systolic blood pressure by mean imputation. Use these filled in values to estimate the mean systolic blood pressure with the corresponding standard error.

(c) Impute the missing values for systolic blood pressure by regression imputation, regressing on gender and age. Write down the regression equation used and the four imputed values obtained. Use these to estimate the mean systolic blood pressure and the associated standard error.

(d) The same as in (c) but now using stochastic regression imputation.

(e) Suppose that hot deck imputation is to be used with strata defined by gender and age (≤ 50 years and > 50 years). Estimate the mean systolic blood pressure with the associated standard error.

Hint: Recall that the standard error of the mean is defined by $\frac{\sigma}{\sqrt{n}}$. Since the population standard deviation is seldom known in practice, we estimate it by the sample standard deviation (denoted by s here), and thus the estimated standard error of the mean is given by $\frac{s}{\sqrt{n}}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. In R you can compute s using the command `sd`.