UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Workshop 1

1. For the situations below, argue whether the missing data are MCAR, MAR, or MNAR.

   (a) A census company mails a census questionnaire to all households in the UK. One-sixth of the households are randomly chosen to receive a long-form questionnaire that asks further questions about income and employment. Assuming all households that receive a questionnaire respond, what is the missing data mechanism for income and employment?

   (b) In a sample survey on employment, a first round of questions is asked of all respondents. To the respondents who are older than 65 years old, further questions on volunteer time are asked. Assume age is known from the first round of questions. What is the missing data mechanism for volunteer time?

   (c) In a sample survey on employment, some people contacted in the survey refuse to give any information and hence are not contacted again. What is a plausible missing mechanism for employment?

   (d) In a laboratory experiment, rats are injected with some bacillus and then assigned either to an active treatment or placebo group. At the end of one week, the concentration of bacillus in the rats is measured. Some rats in the placebo group are dead by the end of that week and so no measurement is recorded. What is a plausible missing mechanism for the concentration of bacillus?

   (e) In a randomised experiment in a factory, different types of blades for cutting steel are tested on steel rods. Occasionally, the machine that transports the rods from the preprocessing area to the cutting area breaks down and no rods are available to be cut. The plan called for ten rods using each type of blade to be cut and then evaluated for smoothness. However, some measurements cannot be made because of the transport system. What is the missing data mechanism for the performance of the blades?

2. Consider data $Y = (Y_1, Y_2, Y_3)$, where the three variables are continuous. The variables Y_1 and Y_3 are fully observed, while some values of Y_2 are missing. Let R be the missingness indicator, taking the value 1 for observed values and 0 for missing values. Because only one variable has missing values, a single missingness indicator suffices. State whether the following mechanisms are MCAR, MAR, or MNAR.

   (a) $\Pr(R = 1 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 Y_1+\beta_2 Y_3}}{1+e^{\beta_0+\beta_1 Y_1+\beta_2 Y_3}}$.

(b) $\Pr(R = 1 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = 0.2$.

(c) $R = 1$ if $Y_1 > y*$, for some known y^*.

(d) $\Pr(R = 1 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 Y_1+\beta_2 Y_2}}{1+e^{\beta_0+\beta_1 Y_1+\beta_2 Y_2}}$.

3. In this exercise we will simulate data and investigate the impact of different mechanisms of missingness (in a similar fashion to what we have done in the lectures). Consider $Y = (Y_1, Y_2, Y_3)$, to be simulated from a standard trivariate normal distribution (i.e., $\mu_1 = \mu_2 = \mu_3 = 0$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$) with correlations $\rho_{1,2}$, $\rho_{1,3}$, and $\rho_{2,3}$ all equal to 0.5. Missingness will then be imposed on Y_2, while Y_1 and Y_3 remain fully observed. Additionally, let R be the missingness indicator, taking the value 1 for observed values and 0 for missing values. In the following consider $n = 500$.

   (a) Impose MCAR by considering $\Pr(R = 0 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = 0.65$. Depict the densities of the complete Y_2 values (as simulated–fully observed), the observed Y_2 values (after imposing MCAR), and of the missing Y_2 values (after imposing MCAR). Comment.

   (b) Impose MAR by considering

   $$\Pr(R = 0 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 Y_3}}{1 + e^{\beta_0+\beta_1 Y_3}}.$$

   Again, depict the densities of the complete Y_2 values (as simulated–fully observed), the observed Y_2 values (after imposing MAR), and of the missing Y_2 values (after imposing MAR). Comment.

   (c) Impose MNAR by considering

   $$\Pr(R = 0 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 Y_1+\beta_2 Y_2}}{1 + e^{\beta_0+\beta_1 Y_1+\beta_2 Y_2}}.$$

   Again, depict the densities of the complete Y_2 values (as simulated–fully observed), the observed Y_2 values (after imposing MNAR), and of the missing Y_2 values (after imposing MNAR). Comment.

   We can play around with the coefficients $\boldsymbol{\beta}$ and check the effect on results. To start with I suggest, $\beta_0 = 1.5$ and $\beta_1 = 3$ for (b), and $\beta_0 = 1.5$, $\beta_1 = 3$, and $\beta_2 = 5$ for (c).