UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Workshop 5

1. When marking the first IDA assignment I read many times: "One advantage of predictive mean matching over stochastic regression imputation is that the former is better suited for imputing heteroscedastic data." However, it was difficult, not to say impossible, to arrive at such conclusion from such a small toy dataset. The goal of this exercise is then to study the performance of predictive mean matching and stochastic regression imputation for imputing heteroscedastic data. To this end, please implement the following steps:

   - Simulate 3000 observations from the following model:

   $$y_i \mid x_i \overset{\text{ind}}{\sim} \mathrm{N}\left(1 + x_i, (3x_i)^2\right), \quad x_i \overset{\text{iid}}{\sim} \mathrm{U}(0,1), \quad i = 1, \ldots, 3000.$$

   - Impose missingness on the response variable by randomly setting 30% of its values to be missing (MCAR mechanism).
   - Using the `mice` package, impute the data using predictive mean matching and stochastic regression imputation. Note that here we are not interested in conducting steps 2 and 3 of multiple imputation. Our goal is to check how well the imputed response values follow the heteroscedastic pattern of the data. Further, and although we are not conducting a formal simulation study, we do not want to be too 'dataset specific', and therefore we set $M = 4$.
   - Inspect the four sets of imputed values. One possible way to do that is to plot a scatterplot of the observed and imputed data. Comment.
   - What about if, instead of stochastic regression imputation, we use its Bayesian counterpart (i.e., if we use `method=norm` in `mice`) to impute the missing values? Do the conclusions remain the same?

2. The National Child Development Study (NCDS) is a continuing longitudinal study that seeks to follow the lives of all those living in Great Britain who were born in one particular week in 1958. The aim of the study is to improve understanding of the factors affecting human development over the whole lifespan. More information can be found here:

   `https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/`

   Here we will be looking at how some early factors affect educational achievement at the age of 23 years old. The file `ncds.Rdata` is available on Learn and contains the following variables:

- `readtest`: childhood reading test score (from 0 to 35) at 7 years old.
- `bsag`: behavioural score (from 0 to 70). Higher values indicate more behavioural problems at the age of 7 years old.
- `sex`: child's sex.
- `care`: in care before 7 years old.
- `soch7`: in social housing before 7 years old.
- `mo_age`: mother's age at child's birth (centred at 23 years old).
- `noqual2`: child has no qualifications at 23 years of age.

Here our goal is to focus on a logistic regression model for the probability of having no educational qualifications at 23 years, `noqual2`, using as covariates `care`, `soch7`, and `mo_age`. Our model of interest is therefore

$$\text{logit}\{\Pr(\text{child has no educational qualifications at 23 years})\} = \beta_0 + \beta_1 \texttt{care} + \beta_2 \texttt{soch7} + \beta_3 \texttt{mo\_age}. \tag{1}$$

All variables in the dataset and, in particular, in the substantive model of interest have missing values. Using multiple imputation, provide the estimates and corresponding 95% confidence intervals of the regression coefficients in (1). There are several steps you need to take: 1) clearly explain your imputation step, 2) check whether the MICE algorithm has converged, and 3) check whether the imputed values are reasonable (in comparison to the corresponding observed values). Note that logistic regression can be easily conducted in `R` with the command `glm` with argument `family="binomial"`.