UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Assignment 3

- To be uploaded to Learn by 4pm, April 13, 2023.
- Location for submission: Gradescope over Learn. **Important**: When uploading your report to Gradescope please tag separately each subquestion (e.g. 1a), 1b), 1c), etc).
- This assignment is worth 40% of your final grade for the course.
- Assignments should be typed (LaTeX, word, etc.).
- Answers to questions should be in full sentences and should provide all necessary details.
- Any output (e.g., graphs, tables) from `R` that you use to answer questions must be included with the assignment. Also, please include your `R` code in the assignment (screenshots of the `R` console are not allowed) or make it available in a public repository (e.g., GitHub).
- The assignment is out of 100 marks.

1. Consider the `nhanes` dataset in `mice`. For more information please type `help(nhanes)` in the `R` console.

   (a) (**2 marks**) What percentage of the cases is incomplete?

   (b) (**4 marks**) Impute the data with `mice` using the defaults with `seed=1`, in step 2 predict `bmi` from `age`, `hyp`, and `chl` by the normal linear regression model, and then pool the results. What are the proportions of variance due to the missing data for each parameter? Which parameters appear to be most affected by the nonresponse?

   (c) (**4 marks**) Repeat the analysis for `seed` $\in \{$`2,3,4,5,6`$\}$. Do the conclusions remain the same?

   (d) (**4 marks**) Repeat the analysis with $M = 100$ with the same seeds. Would you prefer these analyses over those with $M = 5$? Explain why.

2. (**15 marks**) Each of the 100 datasets contained in the file `dataex2.Rdata` was generated in the following way

$$y_i \mid x_i \overset{\text{ind.}}{\sim} \text{N}(\beta_0 + \beta_1 x_i, 1), \quad x_i \overset{\text{iid}}{\sim} \text{Unif}(-1, 1), \quad \beta_0 = 1, \quad \beta_1 = 3,$$

   for $i = 1, \ldots, 100$. Additionally, some of the responses were set to be missing using a MAR mechanism. The goal of this exercise is to study the effect that acknowledging/not

acknowledging parameter uncertainty when performing step 1 of multiple imputation might have on the coverage of the corresponding confidence intervals. Further suppose that the analysis of interest in step 2 is to fit the regression model that was used to generate the data, i.e., a normal linear regression model where the response is y and the covariate is x. With the aid of the `mice` package, calculate the empirical coverage probability of the 95% confidence intervals for β_1 under the following two approaches: stochastic regression imputation and the corresponding bootstrap based version. Comment. For both approaches, please consider $M = 20$ and `seed=1`. **NOTE 1**: In order to calculate the empirical coverage probability, you only need to compute the proportion of the time (over the 100 intervals) that the interval contains the true value of the parameter. **NOTE 2**: The data are stored in an array structure such that, for instance, `dataex2[, , 1]`, corresponds to the first dataset (which has 100 rows and 2 columns, with the first column containing the values of x and the second the values of y).

3. (**9 marks**) Show that for a linear (in the coefficients) regression model, the following two strategies coincide:

   (i) Computing the predicted values (point estimates) from each fitted model in step 2 and then pooling them according to Rubin's rule for point estimates (i.e., averaging the predicted values across the imputed datasets).

   (ii) Pooling the regression coefficients from each fitted model in step 2 using Rubin's rule for point estimates and then computing the predicted values afterwards.

4. The goal of this exercise is to study different ways of using `mice` when the analysis model of interest/substantive model involves an interaction term between incomplete variables. The model used to generate the data (available in `dataex4.Rdata`), which corresponds to our model of interest in step 2, was the following one:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i,$$
$$x_{1i} \overset{\text{iid}}{\sim} \text{N}(0,1), \quad x_{2i} \overset{\text{iid}}{\sim} \text{N}(1.5,1), \quad \varepsilon_i \overset{\text{iid}}{\sim} \text{N}(0,1),$$

   for $i = 1, \ldots, 1000$, $\beta_0 = 1.5$, $\beta_1 = 1$, $\beta_2 = 2$, and $\beta_3 = 1$. Additionally, missingness was imposed on y and x_1 and so the interaction variable $x_1 x_2$ also has missing values, although the missingness in this interaction variable is induced by the missing in the covariate x_1. In the following, please use $M = 50$ and `seed=1`.

   (a) (**6 marks**) By only imputing the y and x_1 variables in step 1, provide the estimates of β_1, β_2, and β_3 along with 95% confidence intervals. Comment. Note that this approach where the interaction variable is left outside the imputation process and calculated afterwards in the analysis model, is known as *Impute, then transform*.

   (b) (**10 marks**) Now, start by calculating the interaction variable in the incomplete data and append it as a variable to your dataset. Then, use *passive imputation* to impute the

interaction variable. Provide the estimates of β_1, β_2, and β_3 along with 95% confidence intervals. Comment.

(c) (**10 marks**) Now that you have already appended the interaction variable to the dataset, impute it as it was *just another variable* (or like any other variable) in the dataset and use this variable for the interaction term in step 2. Provide the estimates of β_1, β_2, and β_3 along with 95% confidence intervals. Comment.

(d) (**6 marks**) What is the obvious conceptual drawback of the *just another variable* approach for imputing interactions?

5. (**30 marks**) The file `NHANES2.Rdata` contains a subset of data from the *National Health and Nutrition Examination Survey* (NHANES), whose goal is to assess the health and nutritional status of adults and children in the United States. The variables in the dataset are the following:

   - `wgt`: weight in kg,
   - `gender`: male vs female,
   - `bili`: bilirubin concentration in mg/dL,
   - `age`: in years,
   - `chol`: total serum cholesterol in mg/dL,
   - `HDL`: High-density lipoprotein cholesterol in mg/dL,
   - `hgt`: height in metres,
   - `educ`: educational status; 5 ordered categories,
   - `race`: 5 unordered categories,
   - `SBP`: systolic blood pressure in mmHg,
   - `hypten`: hypertensive status; binary,
   - `WC`: waist circumference in cm.

   The analysis of interest is the following:

$$\text{wgt} = \beta_0 + \beta_1 \text{gender} + \beta_2 \text{age} + \beta_3 \text{hgt} + \beta_4 \text{WC} + \varepsilon, \quad \varepsilon \sim \text{N}(0, \sigma^2).$$

   Using multiple imputation and conducting all necessary checks, report your findings.

Good work and happy Spring teaching vacation!!