Incomplete Data Analysis

Assignment 1

Callum Abbott

Question 1

Let us suppose that on a (hypothetical) survey there is a question about alcohol consumption and that the variable ALQ records the respondent's answer to the following specific question: "In the past year, have you had at least 12 drinks of any type of alcoholic beverage?". The possible answers are 'Yes' or 'No'. Not all participants respond to this question, that is, the ALQ variable has some missing values. Further, and again hypothetically, suppose that we only have additional data on the gender of the participants in the survey (which is fully observed). For each of the following situations, choose, justifying, the correct answer.

Question (1a)

- **Description**: Suppose that ALQ is MCAR. The probability of ALQ being missing for those with ALQ=Yes is 0.3. What is the probability of ALQ being missing for those with ALQ=No?
- **Answer**: In our hypothetical survey about alcohol consumption two variables are recorded: ALQ and gender. We note that the ALQ variable contains some missing values whilst the gender variable is fully observed. We now suppose that ALQ is MCAR, indicating that the missingness of the ALQ variable is independent of any data. This implies that the probability of the ALQ being missing for those with ALQ=No must **also be 0.3**. If it were any other value, the probability of the data being missing would then have an implied dependence on the ALQ variable rendering the data to **not** be MCAR.

Question (1b)

- **Description**: ALQ being MAR given gender means:
- **Answer**: (ii) The probability of ALQ being missing is independent of the Yes/No value of ALQ after adjusting for gender.

Question (1c)

- **Description**: Suppose again that ALQ is MAR given gender, and that the probability of ALQ being missing for men is 0.1. What is the probability of ALQ being missing for women?
- **Answer**: (iii) It is impossible to conclude from the information given. I believe this because answer (i) of 0.1 would imply that the data would *actually* be MCAR. Meanwhile, answer (ii) seems to indicate that some conservation of probability must be present which is incorrect to assume. This leaves answer (iii).

Question 2

- **Description**: Suppose that a dataset consists of 100 subjects and 10 variables. Each variable contains 10% of missing values. What is the largest possible subsample under a complete case analysis? What is the smallest? Justify.
- **Answer - Largest CCA Subset:** Let us begin with first defining how our data is presented and stored. We can view the data in a table format with 100 rows and 10 columns. The question does not state the total number of missing values in our dataset of $100 \times 10 = 1000$ values, and so let us first assume a scenario where we have 10 missing values (out of 1000) in our dataset. In order to satisfy the criterion that "*each variable (column) contains 10% of missing values*", then we must have 1 missing value in each of our 10 variables. We know that in a complete case analysis that if any one of a subject's 10 variables contain a missing value, then that subject will not be included in the analysis. Therefore, if all our 10 missing values align along a single row in our dataset, then we would only have discard a single subject from our complete case analysis leaving a total of **99 subjects**.
- **Answer - Smallest CCA Subset:** We now imagine the converse situation where we have a much greater number of missing values in our dataset. For argument's sake, let us propose that this number is now 100. This time, in order to satisfy the criterion that "*each variable (column) contains 10% of missing values*", we must have 10 missing values in each of our 10 variables. Now suppose that rows 1-10 were missing for variable 1, rows 11-20 were missing for varible 2, rows 21-30 were missing for variable 3 and so on up to variable 10. This kind of structure in our dataset would result in a missing value being present for every subject in our dataset leaving a total of **0 subjects** under a complete case analysis.