Assessment 2 - Statistical Modelling and Inference for Health

Dr Glen Martin and Dr Matthew Sperrin

This assessment aims to test your understanding of section 2 (causal inference and multiple imputation) of the Statistical Modelling and Inference for Health module. You are required to undertake a data analysis (guided by the indicated questions) of the provided datasets and report your findings. Please format your answers into a report-style and interpret the statistical output. While you are encouraged to provide your R code to document your working, we do **not** expect to simply see R console output copied-and-pasted: it is the interpretation and explanation of your analysis that is needed alongside this. For example, you might like to structure your report with section heading being the questions, and your written answers the body of each section with any code output formatted into tables/figures (with captions).

Assignment settup

This is a data from 11742 individuals affected by Type 2 Diabetes. We want to study whether being treated with metformin (variable 'MetforminTreatment') has an effect on Haemoglobin A1C (variable 'HbA1c'). The variables included in the data set are:

- ID (a sequential row/ID number)
- Gender (F = female, M=Male)
- Smoke (Ex = Ex-Smokers, N= Non Smokers, Y= Smokers)
- Age (years)
- TimeFromDiagnosis (years from time of diagnosis with Type 2 Diabetes)
- HbA1c (Haemoglobin A1C - our primary outcome)
- BMI (body mass index)
- SBP (systolic Blood Pressure)
- Cholesterol (Total Cholesterol)
- Triglycerides
- MetforminTreatment (indicator of whether the patient is being treated with metformin - our primary 'exposure')

After setting your working directory appropriately, load the dataset into R (note the use of tidyverse package):

```
library(tidyverse)
DiabetesData <- read_rds("DiabetesData_Assignment.RData")
```

Part A

1) Summarise the data with appropriate exploratory analysis/plots, including an evaluation of the proportions (and patterns) of missing data. (5 marks)

2) Generate 20 datasets with missing values filled by multiple imputation. To ensure we all get the same 20 datasets, please use set.seed(3333) before imputation. Carefully consider what variables should be included in the imputation models, and comment on the choice of imputation model(s). After running the multiple imputation, check the procedure through appropriate diagnostic information/plots. (15 marks)

3) Considering HbA1c as outcome, suppose all the independent variables in the data set (except for treatment) are potential confounders, perform the following analyses/steps:

a) Extract the "stacked" multiple imputed dataset from question 2 with the following command

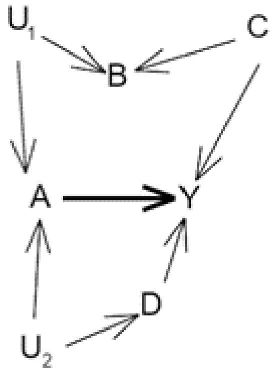
```r
stacked_mi_data <- complete(imp, action = "long", include = TRUE)
#(where imp is the variable storing the mice object from question 2)
```

b) Create a propensity score model on each of the 20 imputed datasets (separately) including all the potential confounders without interactions. Using these models, calculate each individual's propensity score (within each imputed dataset separately) and assign this as a new variable in the stacked imputed dataset called "PS". (10 marks)

c) For each imputed dataset separately, carry out an appropriate analysis to examine the effect of the MetforminTreatment on HbA1c, including the propensity score as an adjustment variable (covaraite) in any models you fit - pool the results across the imputed datasets using Rubin rules. Comment on the use of a propensity score in this way, compared with a model that simply adjusts for all these variables directly, and alternative methods of using the propensity score. (10 marks)

Part B

4) Consider the following DAG



You are interested in estimating the effect of A on Y. The variables U_1 and U_2 have not been measured, all the other variables are available. Assuming the DAG is correct, is it possible to estimate the causal effect of A on Y without bias? Explain your answer. If it possible, state which variables you must adjust for – and if not explain which paths cannot be blocked by observed variables. (10 marks)