Statistical Modelling and Inference for Health – Assessment 2

This assessment aims to test your understanding of section 2 (causal inference and missing data) of the Statistical Modelling and Inference for Health unit. It is worth 50% of your final grade for the whole unit.

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).

The assessment comprises four similarly weighted questions.

* 1. Consider the following DAG:

A picture containing clock, watch, antenna

Description automatically generated

You are interested in estimating the effect of on . The variables and 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 is possible, state which variables you must adjust for – and if not explain which paths cannot be blocked by observed variables. **(10 marks)**

* 1. Suppose your friend is interested in ascertaining whether type 2 diabetes is a cause of lung cancer. They propose to collect a sample of individuals from a large hospital, and then conduct an analysis of the relationship between type 2 diabetes and lung cancer in that sample, accounting for confounders. Explain why sampling from the hospital is a bad idea, and suggest an alternative. **(3 marks).**

The following questions use 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\_Q2 <- read\_rds("DiabetesData\_Assignment\_Q2.RData")

#with similar code for the data indicated for use in question 3 and question 4 below

1. **With the data** **DiabetesData\_Q2.RData**:
   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. 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. **(8 marks) Note: please use set.seed(3333) before imputation**
2. **For this question, use the data** **DiabetesData\_Q3.RData**. This is equivalent to the 20 datasets that you produced in Question 2.
   1. Fit an appropriate regression model to study whether metformin has a causal effect on HbA1c. Clearly state your null and alternative hypothesis. You should combine the results across the datasets in an appropriate way. **(7 marks) Note: you will present and discuss/interpret this model in part b**
   2. Comment on the results in part a). Discuss the assumptions you have made about the missing data to arrive at this result. **(5 marks)**
3. **For this question, use the data** **DiabetesData\_Q4.RData.** This is equivalent to *one of* the 20 datasets that you produced in Question 2.
   1. Create a suitable propensity score model for patients’ receipt of metformin, and check the region of common support. Use 1:1 matching to create a dataset balanced on the propensity score, and use this dataset to establish the causal effect of metformin on HbA1c. **(8 marks) Note: you will present and discuss/interpret the results from the matched analysis in part b**
   2. Comment on the results, any assumptions you have made. Discuss how these results relate to those calculated in Question 3**. (4 marks).**

# Total marks available: 50