Assignment Causal Inference I

Group 8

Introduction

The Framingham Heart Study is a long term cohort study on cardiovascular diseases among people in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1948 and 5,209 persons were initially enrolled in the study. Participants have been examined biennially after inclusion in the study and all participants are continuously followed through regular surveillance for cardiovascular outcomes.

The enclosed dataset is a subset of the data collected as part of the Framingham study and includes laboratory, clinical, questionnaire, and event data on 4,434 participants. Participant clinical data was collected during three examination periods, approximately 6 years apart, from roughly 1956 to 1968. Each participant was followed for a total of 24 years for the outcome of the following events: Angina Pectoris, Myocardial Infarction, cardiovascular disease, Stroke, Hypertension or death.

In this assignment, only data of the first examination period, and the outcomes after 24 years is used. For simplicity all outcomes are considered to be binary (we consider whether or not the event has occurred somewhere in the 24 years, the exact time at which the event occurred is ignored). Note that specific methods were employed to ensure an anonymous dataset that protects patient confidentiality; therefore, this dataset is inappropriate for publication purposes.

The data are in the CVS file: framingham_assignment.csv

Variables in the dataset measured during the first examination period are:

Variable	Description	Units
RANDID	Unique identification number for each participant	
SEX	Participant biological sex	1=Men 2=Women
AGE	Age (years)	
SYSBP	Systolic Blood Pressure (mmHg)	
DIABP	Diastolic Blood Pressure ((mmHg)	
BPMEDS	Use of Anti-hypertensive medication	0=Not currently used 1=Current Use
CURSMOKE	Current cigarette smoking	0=Not current smoker

		1=Current smoker
EDUC	Education	0=High school or below 1=College degree or more
BMI	Body Mass Index, weight in kilograms/height in metres squared	
HEARTRTE	Heart rate in beats/min	

For each participant event data after 24 years is provided. For each type of event, ‘0' indicates the event did not occur, and ‘1' indicates an event did occur.

Variable name	Description
ANGINA	Angina Pectoris
HOSPMI	Hospitalized Myocardial Infarction
STROKE	Stroke
CVD	Cardio Vascular Disease
HYPERTEN	Hypertension. Defined as treated for high blood pressure or having Systolic >= 140 mmHg or Diastolic >= 90mmHg
DEATH	Death from any cause

Assignment

In this assignment, you are asked to estimate the causal effect of cigarette smoking (CURSMOKE) on the risk of Stroke (STROKE)

.

You work on this research question in groups of 3 students during the course. The weekly tasks can be found on Brightspace in the document “assignment tasks per week”. You will write part of a research report with

1. A well formulated research question, in words and using the potential outcome notation.
2. A methods section, with an overview of the causal assumptions made in the analyses, and the methods used to analyse the data.
3. A results section
4. A discussion part

More details can be found in the week by week tasks.

In addition provide:

- The contribution of each author to the report. This can be done by simply stating that "all authors contributed equally to his report" or by letting each author list their own particular contribution to the report.
- The annotated R code used to obtain the results in the report.

Please note the following

1. Your report is primarily graded based on its content. However, layout and writing quality also play a role in the final assessment.
2. Just providing computer output in the report without any explanation does not yield any points.
3. The report (including tables/figures/statement on author contributions, excluding code) may **not exceed 5 pages.**
4. Plagiarism is not allowed and the submitted assignment is screened for plagiarism. The use of generative artificial intelligence to create ready-made content in this assignment is also considered fraud.
5. In case it is stated that all authors contributed equally to the report, one mark for the report will be given. Otherwise each student will get a separate mark based on the indicated division of work and the quality of each student's contributed part.
6. Please state the group number at the first page of the report. We grade reports anonymously, so there is no need to add names, unless the contribution is not the same for all authors.
7. It may happen that we ask you to explain certain parts of the report in an oral meeting.

Grading scheme (slight changes are possible)

	Assigned score		
Category	**0**	**1**	**2**
Research question	Not well formulated and potential outcomes are not or incorrectly used.	Either not well formulated or potential outcomes not or incorrectly used	Well formulated and potential outcomes correctly used
Assumptions with motivation and checks:	None or very incomplete description of assumptions	Some assumptions are listed, some are missing. Some assumptions are not checked or DAG is not very plausible	It is clear which assumptions are used, and assumptions are checked if possible. DAG is plausible
Statistical methods	Statistical methods are poorly described, and/or incorrect or motivation is lacking	Statistical methods are described, but some are incorrect or not well motivated	Correct statistical methods are described with motivation
Results	Results are incorrect or not reproducible from code	Results are incomplete or code is not very clear	Results are correct and reproducible from code
Interpretation	Incorrect interpretation	Interpretation and	Correct interpretation and

and discussion		discussion are superficially	discussion of results
Writing	The report is difficult to read, for example because of lack of structure, errors in writing, wrong tenses, incomplete sentences	There is structure but the report is not easy to read. The style is too informal.	There is a clear distinction between methods, results and discussion. The style is formal. The report is well readable.