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 of the provided dataset and report your findings. **Your report should be a maximum of 10 pages (font size at least 11, standard margins), including any R code and references that you include**.

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 your written answers in the body of each section with any code output formatted into tables/figures (with captions).

The assessment comprises one open-ended question.

A serious disease ‘Heffpox’ is currently affecting a population. Patients diagnosed with the disease have a high chance of death within 7 days. A new treatment ‘milnepan’ has recently been discovered and is thought to reduce this risk of death.

Hundred Acre Hospital is a large hospital that has been treating cases of heffpox using milnepan. The medical director, Dr C Robin, has asked you as data scientist to examine whether milnepan is effective in reducing the risk of death (i.e., prescribing it *causes* a reduction) among patients with heffpox.

A dataset has been extracted for you from the hospital containing data on individuals diagnosed with heffpox, and/or prescribed milnepan (dataset entitled: heffpox\_2324.csv).

Please use this dataset to prepare a report (max 10 pages) for the medical director to address this question. You can assume that the medical director has a good understanding of statistics.

**Additional hints**:

* Don’t forget to conduct an exploratory/descriptive analysis.
* Be sure to be clear on exactly which causal effects that you are estimating.
* Drawing a DAG will help you to identify how to estimate your chosen causal effects.
* The dataset contains missing data and you should address this using techniques covered in the unit.
* You should use at least two different approaches to estimate causal effects and comment on their consistency.
* If you apply multiple imputation you will end up with multiple datasets. If you’re struggling, you can make it simpler if you pick one of the datasets from the imputation to carry forward to the causal analysis, and you will not lose too much credit if you do this. However, students looking to score high marks should propagate the multiple datasets forward to the causal analyses.
* The disease, treatment and hospital are entirely fictitious.

**Dataset information**:

The dataset contains the following variables, as described here.

|  |  |
| --- | --- |
| sex | Sex of the patient, binary (0 = female, 1 = male) |
| age | Age of the patient, integer |
| heffpox | Whether the patient has a diagnosis of heffpox (0 = no, 1 = yes) |
| milnepan | Whether the patient has been prescribed milnepan (0 = no, 1 = yes) |
| death7day | Whether the patient dies within 7 days of admission (0 = no, 1 = yes) |
| bmi | BMI of the patient (weight in kg divided by height in meters squared) |
| icu | Whether the patient is moved to intensive care (which usually happens some days after milnepan is started) (0 = no, 1 = yes) |
| smoking | Whether the patient is a current smoker (0 = no, 1 = yes) |
| diabetes | Whether the patient has a previous diabetes diagnosis (0 = no, 1 = yes). Presence of diabetes is thought to interfere with the action of milnepan. |

**Mark scheme information**:

Approximately half of the marks are allocated for correct and appropriate analysis, with the other half being allocated for clear descriptions and justifications of the analyses, their interpretation, discussion of the implications, and presentation. It is particularly important that you clearly explain what it is you are estimating, your assumptions, and conclusions. Simply presenting code and output with no discussion and explanation is inadequate and will result in a fail.

The report must be a maximum of 10 pages in length. If your report is longer than this, any context beyond the 10 pages will not be included for marking. You should not do things like reducing font size and figure size, reducing margins to get within the page limit – instead, think about how your analysis can be made more concise.

The below table gives an indication of how your final mark corresponds to your report:

|  |  |
| --- | --- |
| **Mark** | **Descriptor** |
| 0-19% | Analysis makes little or no attempt to address the brief, is generally inaccurate and is hard to follow. Conclusions are either nonexistent or incorrect. |
| 20-29% | Analysis addresses the brief in a limited way but is generally inaccurate. Report is generally incoherent. Conclusions often incorrect, and no attempt is made to discuss assumptions of analysis. |
| 30-39% | Analysis addresses only parts of the brief and is often inaccurate. Report is generally incoherent. Conclusions often incorrect, and no attempt is made to discuss assumptions of analysis. |
| 40-49% | A partly correct analysis that has addressed at least some of the brief. Report lacks coherence on the whole. Conclusions are sometimes incorrect, and little or no attempt is made to discuss assumptions of analyses. |
| 50-59% | A mostly correct analysis that has addressed the brief. Report is coherent but may be hard to follow in places. Conclusions and interpretation generally correct but superficial. Some limited effort is made to discuss assumptions of the analysis. |
| 60-69% | Analyses are correct besides minor errors and typos. Report is coherent throughout. Conclusions and interpretation are correct but lack depth. Reasonable discussion of assumptions and limitations of analysis. |
| 70-79% | Analyses are correct, besides minor errors and typos, and address the brief in considerable detail. Conclusions and interpretation are correct and detailed. Comprehensive discussion of assumptions and limitations. |
| 80-89% | Analyses are correct, address and often go beyond the brief, demonstrating independent thought and judgement. Conclusions and interpretation are correct and detailed. Comprehensive discussion of assumptions and limitations. |
| 90-100% | Marks in this category reflect a report of publishable quality with advanced insights and going well beyond the brief. Reports may demonstrate novel and innovative approaches to address the brief, and interpretations and conclusions will demonstrate a deep understanding and judgement. |

This assessment tests the following intended learning outcomes of the unit:

|  |  |
| --- | --- |
| **Category of outcome** | *Students should be able to:* |
| A. Knowledge and understanding | LO1: Explain and discuss modelling and causal inference techniques and appraise their application in healthcare  LO2: Appraise the strengths and weakness of modelling methods  LO3: Discuss the challenge of causal inference and the strong assumptions required. |
| B. Intellectual skills | LO5: Assess the effectiveness and fitness for purpose of a modelling tool or technique  LO6: Apply modelling techniques and methods to healthcare data  LO7: Interpret analytical results |
| C. Practical skills | LO8: Design and write scripts to implement statistical /mathematical methods to analyse health data |
| D. Transferable skills and personal qualities | LO9: Communicate and write analytical methods based on completed work and available literature in this area  LO10: Develop problem solving skills  LO11: Demonstrate a critical understanding of technical descriptions of statistical/mathematical analysis methods |