HDAT9700 Statistical Modelling II

Assessment 2 Instructions

Mark Hanly

Overview

For your final assessment you are required to prepare a short analysis report applying one of the statistical techniques covered in the course. The report should be presented in IMRD (Introduction/Methods/Results/Discussion) format and should be approximately 1500 words.

Topic options

You can base your report on one of the following three options:

Causal analysis via matching Exploring the causal effect of young maternal age on perinatal outcomes

Interrupted time series Evaluating the effect of media coverage following publication of new drug safety research on dispensing of oral contraceptives using time series data

Longitudinal multilevel modelling Modelling children's weight change between 6 - 24 months

A dataset is supplied for each option. A more detailed description of each option is provided at the end of this document.

Report structure

Your report should be presented in IMRD format and be approximately 1500 words long. The word count does not include tables, figures, footnotes, references, appendices, or the student declaration. You should include up to five tables/figures in the main body of the report, with supplementary tables/figures included in the appendix.

Below is in an example of how you might lay out your report:

1. Student declaration
2. Introduction (200-300 words approx.)
3. Methods (550-650 words approx.)
4. Results (350-450 words approx.)
5. Discussion (200-300 words approx.)
6. References (As appropriate)
7. Appendix (As appropriate)
8. Student declaration

Introduction

200-300 words approx.

The introduction section should provide some brief background information on the research area and motivate why the research question is of scientific interest. Because you have been presented with a research topic you do not need to make

this section very detailed, you should simply demonstrate that you have understood the relevance of the research question, and make the connections between the research question, the available data and the analytic approach.

Methods

550-650 words approx.

In the methods section you should describe in detail the analyses that were undertaken, and why. It is important that you demonstrate that you understand the rationale for applying a particular statistical technique and/or any alternative model specifications/model diagnostics. You should demonstrate your understanding of the statistical methods with appropriate references to the literature.

Results

350-450 words approx.

In the results section you should begin with a descriptive overview of the dataset you have used, including summaries of any key variables. Follow this with a summary of your analysis results using appropriate tables and figures, as well as describing your main results in text. Appropriate model diagnostics and/or alternative model specifications should be included but can be presented in the appendix if they are lengthy.

Discussion

200-300 words approx.

In the discussion you should summarise your main findings, interpret your results in more detail, discuss implications, and acknowledge strengths and limitations of the analysis, with an emphasis on the statistical methods applied.

References

Your report should include 5-10 appropriate references to the literature, with a full list of references provided in the references section. You can use your preferred referencing style. More information on academic referencing can be found on the UNSW website: https://student.unsw.edu.au/referencing

Appendix

The appendix should include any supplementary material that does not fit in the body of the report, for example additional descriptive tables or model diagnostics plots.

Student declaration

All assessments should include a copy of the student declaration regarding plagiarism, student academic misconduct and proper back-up of your assessment. The text of the declaration is provided below. You should copy this into your report and tick the boxes to indicate your agreement with the statements.

> I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere or previously, or produced independently of this course (e.g. for a third party such as your place of employment) and acknowledge that the assessor of this item may, for the purpose of assessing this item: (i) Reproduce this assessment item and provide a copy to another member of the

> University; and/or (ii) Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).
> - [ ] I understand and agree I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.
> - [ ] I understand and agree I have a backup copy of the assessment.
> - [ ] I understand and agree

Accessing and submitting your report

The data for each assessment option, and any additional hints or instructions, will be made available in a GitHub repository. You can clone this repository through a hyperlink provided by your course convenor. To submit your assessment files, add and commit them to the repo, and then push to GitHub. You are welcome to explore all three assessment options before deciding which option to choose. You may submit your assessment as either:

1. A Microsoft Word document with the corresponding R code provided separately
2. an R Markdown file, interweaving the text of your report, statistical code and results.

Statistical code

You should carry out all data preparation and statistical analyses using R. Include all R code you use to prepare and analyse the data as part of your submission. Following best practice in reproducibility, it should be straightforward for the marker to reproduce your entire analysis using only the raw data and the code you submit. Your code should be clearly commented, to assist in this reproducibility, and to demonstrate your understanding of the analyses being performed.

Marking rubric

Your assessment will be marked according to the marking rubric on the following page. Four criteria are assessed: • Conceptual understanding (30%) • Application of statistical method (30%) • Interpretation (30%) • Presentation (10%) 6. Plagiarism

While you are welcome to discuss the assessment with your classmates, the final written report and statistical code must be your own work. References to other scholarly work should be properly cited. **Direct plagiarism will not be accepted.** Past incidences of plagiarism have resulted in students retaking the assessment. You can read academic integrity at UNSW and examples of plagiarism here: https://www.student.unsw.edu.au/plagiarism

Submission deadline and late submissions

Assessment 2 is due by **9am on Monday 15 August 2022**. A penalty of 5% per day will apply to late submissions. To allow marking to take place in time for the university grade submission deadlines, late assessments cannot be accepted after Thursday 18th August (barring special circumstances). Please notify the course convenors as soon as possible in the case of extenuating circumstances that prevent submission by this date.

Option 1: Causal inference via matching

Overview

Young maternal age is associated with adverse pregnancy outcomes, including low birthweight and preterm birth. As a result, in some contexts “teenage pregnancy” has been framed as a public health problem. However, many academics and health professionals have argued that young maternal age doesn’t cause adverse outcomes, but rather socio-economic disadvantage is the primary factor underlying different perinatal outcomes experienced by younger and older mothers. (cf. Lawlor and Shaw (2002) “Too much too young? Teenage pregnancy is not a public health problem”, International Journal of Epidemiology (31) p552-554, and the subsequent counterpoint articles, for more on this debate).

In this assignment you will use data on maternal and family demographics and child birth outcomes, to estimate the causal effect of young maternal age on the risk of preterm birth. Use the MatchIt package in R to identify a matched “treatment” and control group, who are as similar as possible in terms of appropriate pre-treatment variables. A DAG should be used to guide variable selection and highlight limitations of the available data.

Research question

> What is the causal effect of young maternal age on the risk of preterm birth?

Data

The data for this assessment are drawn from the National Children’s Study Archive. The National Children’s Study (NCS) collected birth and early childhood data on more than 5,000 children and their families in the USA from 2009-2014. The NCS teaching database contains base child-level data, as well as several topic-specific modules.

You are provided with a dataset named **ncs_child_mom_matching.csv**. This file contains information on 3,848 singleton births including mother and child demographics, birth outcomes and area-level statistics. The data are drawn from the NCS base Child dataset, and the Mother’s Pregnancy Health module. Some data preparation has already taken place, in particular, the aforementioned Child and Mother datasets have been merged, and variables with high proportions of missing data have been removed.

The following resources will be useful to help understand the data: * Information on the National Children’s Study Archive * NCS study description and guide * Codebook for the base Child data * Codebook for the Mother’s Pregnancy Health module Hints * Young maternal age is usually defined as maternal age less than or equal to 19 years. * Preterm birth is usually defined as gestational age less than or equal to 37 weeks. * Use the MatchIt package in R to explore different matching approaches. * Assess balance between the treated and matched control groups using numeric and graphical summaries. * Not all variables are necessarily appropriate for matching and not all variables you would like to match on will be available. Use DAGs to guide variable selection and highlight limitations. * Remember to apply the matching weights when analysing matched data.

Option 2: Interrupted time series analysis

Overview

Oral contraceptives (“birth control pills”) are taken by women to prevent pregnancy and to treat other conditions such as endometriosis or polycystic ovary syndrome. There are two main types of oral contraceptive pills, the most common being the combined pill, which contains a combination of the hormones oestrogen and progestogen in various formulations and

doses. The second type is the progestogen-only pill (also called the “mini pill”) that contains only progestogen.

Although they are effective at reducing the occurrence of unwanted pregnancy, the combined pill increases the risk of blood clot formation, such as deep vein thrombosis, pulmonary embolism and stroke. A paper published in the BMJ on 26 May 2015 was the first to quantify the risk associated with different oestrogen/progestogen combinations and received substantial media attention worldwide (Vinogradova et al. Use of combined oral contraceptives and risk of venous thromboembolism: nested case-control studies using the QResearch and CPRD databases. BMJ 2015;350:h2135). Note that there is no increased risk of blood clots in women using the progestogen-only pill.

In this assessment you will use time series data to explore the impact of the media attention surrounding the publication of this study on the PBS-subsidised dispensing of combined and progestogen-only oral contraceptives. In Australia, combined contraceptive pills containing levonorgestrel/ethinylestradiol, norethisterone/ethinylestradiol and norethisterone/mestranol are subsidised through the PBS. Other combined pill formulations are not subsidised, such as those containing drosperinone, cyproterone and desogestrel and are not captured in the data. Several progestogen-only mini pills are also subsidised, including levonorgestrel and norethisterone.

Research question

> What was the impact of the media attention around the increased risk of thrombosis associated with oral contraceptive use on dispensing of combined oral contraceptives?

Data

Data are provided in the file contraceptives.csv. This file includes monthly counts (per 10,000 women of reproductive age) of PBS-subsidised dispensings of combined oral contraceptives (“combined”) and progestogen-only contraceptives (“mini”) between Jan 2013 and Dec 2016. The peak of the media attention was in the last week of May 2015. Hints * Use appropriate visualisations to explore the raw data. * Explore seasonal effects, autocorrelation, and delayed impacts with statistics and tests. * Create appropriate variables to model changes after the intervention. * Use an appropriate model to explore whether there was a change in combined oral contraceptive dispensing following the media attention. * Plot the predicted counterfactual to visualise any change in dispensing following the media attention. * Use the “control” series (dispensing of progestogen-only oral contraceptives) and comment on the findings in relation to those from the combined pill series.

Option 3: Longitudinal multilevel modelling

Overview

Children’s growth is frequently monitored during their infancy and understanding growth trajectories is important for identification of chronic disorders or other health issues, reassuring parents and providing information on the health of a nation’s children.

In this assignment you will use data including repeated measurements of children’s weight to model children’s weight change over time. The data also include information on other infant, environmental and maternal factors, and you will also explore which of them could further improve the model for children’s weight gain. You should use the nlme package in R for carrying out your longitudinal multilevel modelling.

Research question

> What is the optimal longitudinal multilevel model for modelling children's weight change?

Data

The data for this assessment are drawn from the National Children's Study Archive. The National Children's Study (NCS) collected birth and early childhood data on more than 5,000 children and their families in the USA from 2009-2014. The NCS teaching database contains base child-level data, as well as several topic-specific modules.

You are provided with a dataset named **ncs_child_repeated.csv**. This file contains information on child, mother and household demographics for 4,531 kids, with 1-4 weight observations for each child (12,654 observations overall). The data are drawn from the NCS base Child dataset, and the Child health module, including children's weight measurements at 6, 12, 18 and 24 months after their birth. Some data preparation has already taken place, in particular, the above files have been merged, only singleton births and the oldest siblings have been included (i.e. only one child per mother), and variables with high proportions of missing data have been removed.

The following resources will be useful to help understand the data: * Information on the National Children's Study Archive * NCS study description and guide * Codebook for the base Child data * Codebook for the Child Health module

Hints

- The outcome weight (`visit_wt`) can be modelled as a function of continuous age (`child_age`) or ordinal visit timing (`visit`).
- Conduct appropriate EDA to understand the relationship between weight and age.
- Use the `nlme` package for carrying out your multilevel modelling, starting with simple models and progressively exploring more complex structures.

- There is only one child per mother, and thus you do not need to include mother as an additional level in your multilevel modelling.
- Explore alternative functions of time to improve model fit.
- Explore other covariates that help to explain variation in children's weight.
- Some child health variables (e.g. `gastro`, `diarrhoea`, `ear_infection`) were not measured on every visit occasion, so if these are included you will need to think about how to handle that.

Marking rubric

HDAT9700 Statistical Modelling II, Assessment 2 Marking Rubric

Criteria	Weight	Developed (HD/DN)	Developing (CR/PS)	Developed (FL)
Conceptual understanding	30%	Demonstrates a deep understanding of the relevance of the statistical technique, it's appropriateness to the research question at hand, its's benefits over alternative approaches, but also its limitations. Incorporates content beyond what was covered in the	Demonstrates an understanding of the statistical technique, but to a limited level, without delving into appropriate additional readings signposted in the tutorial material.	Demonstrates limited understanding of the statistical method, its suitability to the research design, advantages or limitations.

Criteria	Weight	Developed (HD/DN)	Developing (CR/PS)	Developed (FL)
		tutorial, such as the pre-reading and additional references cited in the course notes, or other scholarly work.		
Application of statistical method	30%	Analysis goes beyond the basics covered in the face-to-face tutorial. Statistical method is correctly applied, selection of variables is justified where appropriate, alternative model fits are explored, where appropriate; suitable diagnostic tests are performed, presented and correctly interpreted. The rationale for carrying out models/diagnostics is clearly understood and communicated. All code has a clear purpose and runs without errors.	Analysis is largely correct but does not go beyond replication of what was covered in the tutorial. Limited model specifications and diagnostics are explored or are presented without demonstrating an understanding of why they are performed. Some code has unclear purpose or produces an error.	Inappropriate statistical method is performed, or correct method is performed but with major errors invalidating the results; large sections of code does not run or has unclear purpose; appropriate diagnostic tests are not performed, are misinterpreted.
Interpretation	30%	All relevant statistical output is clearly presented. Output is interpreted correctly and thoroughly. Implications follow clearly from the analysis, are nuanced, and demonstrate appreciation for the strengths and limitations of the analysis.	Analysis output is presented adequately but lacks some important features. Interpretation lacks nuance or has minor inaccuracies or misinterpretations. Implications, strengths and limitations are discussed at a superficial level.	Correct output is not presented (or all output is presented indiscriminately, without interpretation or highlighting the important statistics). Interpretation is incorrect or lacking entirely; implications are absent or do not follow from the analysis. Strengths and limitations are inaccurate or absent.
Presentation	10%	Report is well-written and professionally formatted; code is rigorously documented; tables and figures are presented to a publication-quality, with clear titles, labelling and legends; references are appropriate and presented consistently.	Report is adequately presented, with only a small number of typographical or grammatical errors. Tables and Figures are interpretable but lack some details to reach publication standard, such as inadequate labelling, legends or scaling. References are present but have minor errors. Code commenting is present but not detailed.	Report is poorly written and presented, with typos and grammatical errors. Tables/figures are not presented or are difficult to interpret, with insufficient headings/ labels/ legends etc; code is poorly commented; references are absent; or not presented in a consistent academic format.

Note:

HD=Higher Distinction (85-100%); DN=Distinction (75-84%); CR=Credit (65-74%); PS=Pass (50-64%); FL=Fail (0-50%)