MSc in Statistical Science
Generalised Linear Models – Assessed Practical

Week 8, MT 2025

- This practical sheet contains two sections. **Write a report on the Exercise in Section 2 only.**
- **The report has a word limit of 2000 words**. This word limit is on the main body of the report. Equations, tables, figures, captions and appendices to your report do not contribute to the word count. Include your R-code in an appendix.
- **You should use your anonymous practical ID (of the form P123 and not your name)** for the cover page of the report.
- **You should submit your report via the Inspera system**.
- **The hand-in deadline is 12 noon Wednesday 10 December 2025** .

Any queries you have about the exercise in Section 1 may be directed to the lecturer during the practical session. The lecturer will not answer questions regarding the exercise in Section 2, with the sole exception of questions relating to a limited number of programming issues.

Use of Generative AI

The department's policy on the use of generative AI in assessed work is appended to this handout.

- You should ensure that you have read and understood this policy.
- In your submitted work, you must abide by the policy and include a "Declaration of AI use" section at the start of your report.
- The policy document includes a template for your declaration and an example declaration.

The policy is also available at:

`https://canvas.ox.ac.uk/courses/295123/files/8460637?module_item_id=3075072`

1 Exercise for practice, NOT ASSESSED

The dataset `bw.csv` gives details of 189 babies and mothers, focusing on low birth weight. The dataset contains information on:

- `low:` birth weight status, 1 = birth weight less than 2.5 kg, 0 otherwise
- `age:` mother's age in years
- `mwt:` mother's pre-pregnancy weight in pounds
- `race:` mother's self-reported ethnic group (variable named `race` in the original dataset), 1 = White, 2 = Black, 3 = any other ethnic background
- `smoke:` 1 if smoked during pregnancy, 0 otherwise
- `ptlp1:` 0 if no previous premature labours, 1 otherwise
- `ht:` 1 if mother has history of hypertension, 0 otherwise

(a) Produce summary statistics and some exploratory plots of the data, both univariate and bivariate, examining the relationships between low birth weight status and the other variables.

(b) Which GLM do you specify to analyse how the incidence of low birth weight depends on the other variables? Motivate your choice.

(c) Carry out model selection (ignoring any interaction terms).

(d) Assess the quality of the model fit using suitable methods.

(e) Interpret your findings fully.

(f) Compute an estimate of the average marginal effect for `mwt`.

2 ASSESSED EXERCISE

The data in `airpollution.csv` relate to a survey on air pollution in a city in South Asia. Each row corresponds to the answers of one individual. The variables are:

- `respd`: binary variable (Yes/No) indicating whether the individual experienced any respiratory difficulties in the last 24 hours
- `age`: age of the individual in years (25, 35, 45, 55) representing four 10-year age bands
- `gender`: gender of the individual (Female/Male)
- `activity`: binary variable (Yes/No) indicating whether the individual had any physical activity in the last 24 hours
- `airpollution`: binary variable (Yes/No), indicating if there was an event of extreme air pollution in the last 24 hours

The data can be loaded using
`dat <- read.csv("airpollution.csv", stringsAsFactors = TRUE).`

Exercise:
Investigate and write a report on how the probability of experiencing respiratory difficulties depends on the other variables. The main goals are to obtain a suitable interpretable model and to give a full interpretation of that model.

(a) [Exploratory Data Analysis, 6 marks] Provide simple plots and/or tables that examine the distribution of the different variables and examine the relationships between respiratory difficulties and the available explanatory variables. Comment on and summarise your findings.

(b) [Modelling, 5 marks] Model the relation between `respd` and the other variables that are available using a Bernoulli GLM with canonical link function. Using appropriate tests, carry out model selection to examine the relationship between the possible explanatory variables and the probability of experiencing respiratory difficulties. When considering interactions, restrict attention to interactions between `airpollution` and each of the other explanatory variables.

(c) [Model Diagnostics, 4 marks] Assess the quality of the model fit using suitable methods.

(d) [Interpretation, 6 marks] Carefully interpret the estimation results of your final model as obtained in (b), with all observations included. Present and interpret the corresponding odds ratios (multiplicative effects on the odds) for each variable in your final model, with 95% confidence intervals.

(e) [Prediction, 4 marks] Provide an estimate for the probability of a female, aged 45, who has had some physical activity during an extreme air pollution event, to experience respiratory difficulties. Give also a confidence interval. Provide the same quantities for a male, 85, who has had no physical activity during an extreme air pollution event. Comment on your findings and discuss the reliability of such predictions.

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

Policy for students on the use of generative artificial intelligence in summatively assessed work

Scope of this policy

In the Department of Statistics, this policy applies to: undergraduate SB1 practicals, Part C and OMMS dissertations, MSc Statistical Science SB1 and SB2 practicals, and MSc Statistical Science dissertations.

For submissions for the MSc Statistical Programming assignment, the current default policy is that unauthorised use of AI is prohibited.

The policy below is the same as that for similar courses in the Mathematical Institute.

Broader context of this policy

The department encourages experimentation with, and the confident use of, generative artificial intelligence (Gen AI) tools in your studies. Gen AI significantly extends the established tool kit such as search engines, spell-checkers, or debuggers, and opens up new opportunities in collating material, improving your writing, or generating code. However, these wide-ranging and rapidly expanding capabilities require careful consideration in relation to originality, attribution, and good academic practice.

The present document specifies the permitted use of generative AI tools in assessed work and the requirements for declaration of such use, as per the university's AI policy from July 2025[1]. The consequences of unauthorized use through the University's procedure for plagiarism range from mark deductions in minor cases of bad academic practice to expulsion from the university in extreme cases.

You are bound by the university's declaration of authorship for all taught submissions, in which you pledge, among others, "that submitted work is entirely [your] own, except where otherwise indicated"[2].

[1] https://academic.admin.ox.ac.uk/ai-use-in-summative-assessment
[2] https://www.ox.ac.uk/students/academic/exams/open-book/honour-code

General considerations for the use of Gen AI

Gen AI can assist, but not replace, your own critical engagement with the subject. It is therefore imperative to carefully examine any Gen AI output you build on, and be aware of the limitations and pitfalls of the tools you use. These include, but are by no means limited to: reproducing information the model was trained on, potentially leading to classical plagiarism and copyright infringement; 'hallucinations', i.e., presenting false information as fact, either through training on erroneous data or incorrect synthesis of correct data, such as made-up references; producing biased or offensive output; predictive output that does not meet the standard of novelty and/or originality expected from the research work in question; code with incorrect syntax or which gives unintended results. Ultimately, you, the student, take full responsibility for the submitted work.

It is also worth pointing out that in mathematical writing you are not assessed on the correct or creative use of English, but rather the clarity and logical structure of your exposition. If you use a Gen AI tool to assist you in your writing (e.g., to eliminate grammatical mistakes), it is essential that you scrutinise the result and edit it such that it captures precisely what you want to say.

Permissible and impermissible use

As a broad principle, if it would be inappropriate to ask someone else to do a certain part of the work for you, it will also be inadmissible to use AI for that task. In the same way as it would be inappropriate to use existing mathematical results and software without acknowledgment, you are required to acknowledge the use of Gen AI to produce code or mathematical derivations.

Examples of generally **permissible** use of AI:

- searching for literature on a specific topic, provided you verify the relevance, correctness and unbiasedness of the references, and the discussion of the sources is your own;
- soliciting feedback on written work to improve clarity or reduce length;
- generating ideas for research questions, extensions or examples;
- assistance in writing code[3], including generation of new code and translation of existing code into a different programming language; or improving the efficiency of code;
- non-substantive alterations to figures that improve clarity and readability;
- formatting of lists or bibliographies;
- improving your grammar.

[3] except in specific coding assignments, or if explicitly not allowed to use existing code.

In all cases, Gen AI use needs to be declared as outlined above. All Gen AI output has to be checked, and any errors may have an impact on the final mark, even where properly cited.

Examples of generally **impermissible** use of AI:

- substantive original writing by Gen AI, e.g., for an introduction or conclusion section, including verbatim or closely paraphrased use of Gen AI for a literature review;
- generating sections of improved versions of your original writing beyond spelling and grammar, e.g., that bring in domain knowledge or improve a logical argument;
- any use of AI generated code without declaration;
- production of plots by Gen AI, as this may obfuscate the data generating algorithm or the data source, including alteration of data in a diagram;
- direct use of Gen AI interpretations of mathematics or data.

Always follow the departmental[4] and the university's[5] information security policies and the university’s data protection policy[6] when interacting with Gen AI.[7] In particular, do not enter private or confidential data into third-party AI tools as they may be stored or passed on to a wider audience, in their original or a processed form, for instance, by being used to train the Gen AI tool. An example of such data would be internal data provided by an industrial partner or data obtained under a licensing agreement.

Declaration of AI use

The precise use of AI tools has to be declared in the same way that existing literature or available software has to be referenced adequately and should include the model and version of the tool used (e.g., GPT-4.1). This declaration has to include how the Gen AI tool was interacted with. While it is not generally expected that you give a complete list of individual prompts used, you need to make clear, for instance, whether a piece of software was constructed through an iterative process with Gen AI by building individual blocks, or fully AI generated.

Template for declaration of AI use:

In creating this work, I have used AI tools for the following tasks:

Literature search and ideation:

Spelling and grammar:

[4] https://www.maths.ox.ac.uk/members/it/it-notices-policies/information-security-policy
[5] https://www.infosec.ox.ac.uk/guidance-policy
[6] https://compliance.admin.ox.ac.uk/data-protection-policy
[7] https://infosec.ox.ac.uk/use-generative-ai-services-such-as-chatgpt-safely

Improvements to writing:

Coding:

Plotting, formatting etc:

Example declaration (NB: not to be seen as an endorsement for the tools mentioned):

In creating this work, I have used AI tools for the following tasks:

***Literature search and ideation:** e.g. "I used Perplexity (free version Oct 25) through the Comet browser to find literature on models for non-Newtonian fluids, and cross-checked this against Google Scholar. I asked Claude 4.0 to summarize the main contributions in those articles. I then skim-read the most promising papers to check relevance for my project and discussed those in my own words in my literature review."*

***Spelling and grammar:** e.g. "I used Grammarly within Overleaf to detect spelling and grammar errors throughout the essay."*

***Improvements to writing:** e.g. "I used ChatGPT 5 Edu to get feedback on the clarity of my explanation of the main contributions. Sentences were remodeled occasionally on a sentence-by-sentence basis, without major structural changes or without changing the meaning conveyed for the paragraph. For the final version, I asked for suggestions to reduce the length of the dissertation to the paper limit. Following the suggestions, I moved material in an appendix and tightened the discussion of the numerical output."*

***Coding:** e.g. "I used Gemini 2.5 for code diagnostics—to assist in identifying and rectifying minor structural or syntax-related errors in the implementation of the numerical scheme."*

***Plotting, formatting etc:** e.g. "I used ChatGPT 5 Edu to create a tikz template for illustrating the neural network in Figure 5.1 in latex; to bring the bibtex bibliography into a unified format; as a search engine for plotting commands in python."*

Taking text or mathematics verbatim from generative AI is strongly discouraged, as it is not reproducible, and cannot be assigned authorship or cited as a permanent source. Any such verbatim passages need to be flagged up explicitly (e.g., in quotation marks) with citation in the style of a "personal communication".