Take Home Assessment 2025

Ioanna Manolopoulou, Niloufar Abourashchi and Nuru Giritli

2025-03-20

Rules for the take home assessment

- This assessment is classified as Coursework as defined in the UCL Student Regulations for Exams and Assessments (https://www.ucl.ac.uk/academic-manual/chapters/chapter-4-assessment-framework-taught-programmes/student-regulations-exams-and-assessments). It contributes 75% to the overall mark for this module.
- The release date for this assessment is 10:00 (UK time) on Friday, 21 March 2025.
- **The deadline for submission is 12:00 (noon) UK time Friday, 2nd May 2025**
- Your group will submit, via the Submit your take home assessment (https://moodle.ucl.ac.uk/mod/assign/view.php?id=5643200) Moodle page, one PDF file containing your report and one R/Rmd program file.
- Each group member must click their respective “Submit” buttons in order for the group’s submission to be successful and final. By ticking the submission declaration box in Moodle you are agreeing to the following declaration:

> ***Declaration:*** *I am aware of the UCL Statistical Science Department’s regulations on academic misconduct for assessed coursework. I have read the guidelines in the student handbook and understand what constitutes academic misconduct. I hereby affirm that the work my group is submitting for this in-course assessment is entirely our own.*

- Parts of your submission may be scanned using similarity detection software. If any breach of the assessment regulations is suspected, it will be investigated in accordance with UCL’s Student Academic Misconduct Procedure (https://www.ucl.ac.uk/academic-manual/chapters/chapter-6-student-casework-framework/section-9-student-academic-misconduct-procedure).
- You will work together within your group and the usual plagiarism and collusion regulations do not apply to this form of interaction. However, they do apply to collusion with other groups or plagiarism of work from other groups or from other sources.
- The best source of support to complete this assessment will be through the dedicated drop-in hours with the instructors as well as the Moodle forum.
- Use of Artificial Intellignce (AI) is permitted for this assignment and falls under Category 2 of UCL’s guidance on use of AI in assessments (https://www.ucl.ac.uk/teaching-learning/generative-ai-hub/using-ai-tools-assessment#:~:text=Category%201%3A%20AI%20tools%20cannot,of%20AI%20and%20referencing%20AI). This means that you can use generative AI in an assistive role, however your report should be your own work, otherwise it will be classified as plagiarism.

- To facilitate anonymous marking, you should not write your name anywhere on your work, including in file names or file descriptions requested as part of the submission process.
- Non-submission (in the absence of any valid extenuating circumstances) will mean that your mark for this component is recorded as 0.00% and you will be deemed to have made an attempt.
- Any plagiarism will normally result in zero marks for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in Departmental Student Handbooks. The relevant exerpt (https://moodle.ucl.ac.uk/mod/url/view.php?id=5852915) from the Statistical Science handbook is also posted on Moodle. You may find it useful to take the academic micsonduct quiz (https://moodle.ucl.ac.uk/mod/quiz/view.php?id=5864881) on the STAT0004 moodle page.
- There are strict, non-negotiable penalties for late submission (https://www.ucl.ac.uk/academic-manual/chapters/chapter-4-assessment-framework-taught-programmes/section-3-module-assessment#3.12), which for coursework are as follows. - Up to 2 working days late: deduction of 10 percentage points, but no lower than the pass mark. - 2-5 working days late: capped at the pass mark. - More than 5 working days late: mark of 1.00%.
- Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation. Check guidelines here (https://www.ucl.ac.uk/academic-manual/chapters/chapter-2-student-support-framework/2-short-term-illness-and-other-extenuating). Penalties are set out in the latest editions of the Statistical Science Department student handbooks, available from the departmental web pages. Extensions to the submission deadline can only be granted where a student from the group has been issued with a SoRA (https://www.ucl.ac.uk/students/support-and-wellbeing/disability-support/reasonable-adjustments-your-assessments) (note: SoRA extensions for groupwork have to be requested by email to us) or has made a valid claim for extenuating circumstances (https://www.ucl.ac.uk/academic-manual/chapters/chapter-2-student-support-framework/2-short-term-illness-and-other-extenuating). The standard extension length for this assessment type is one week.
- Extenuating circumstances are handled by your parent department and all claims should be submitted via Portico (https://www.ucl.ac.uk/academic-manual/chapters/chapter-2-student-support-framework/2-short-term-illness-and-other-extenuating-1). Depending on the nature and severity of the circumstances, an alternative type of mitigation to a deadline extension may be considered more suitable.
- Failure to submit this in-course assessment may mean that your overall examination mark is recorded as “non-complete”, i.e., you will not obtain a pass for the course.
- All members of a group will be awarded the same mark for the assignment, unless a member is flagged to us as inactive before submission.
- We may ask you as a group to come and discuss your output with us.
- You will receive, via Moodle, feedback on your work and a *provisional grade — grades are provisional until confirmed by the Statistics Examiners’ Meeting in June 2025*.

Instructions and marking scheme

Data

The data for the Take Home Assessment, available under the **Data Files** folder in Moodle, describe the citation relationships of papers submitted to **arXiv.org** in the category of *High Energy Physics – Phenomenology* between **1993 and 2001**. You are given two files: `citations.txt` and `papers.txt` .

1. **papers.txt**

Each row in `papers.txt` has two comma-separated values:

i. **Paper ID**: You can access the papers on arXiv by appending `arxiv.org/abs/hep-ph/` before the paper ID. If the paper ID starts with **11**, it is a cross-listed paper from the *High Energy Physics – Theory* section, and you can access it by appending `arxiv.org/abs/hep-th/` to the last 7 digits of the paper ID.
ii. **Submission Date**: The date the paper was submitted to arXiv.org.

2. **citations.txt**

Each row in `citations.txt` has two comma-separated values:

i. **FromID**: The ID of the paper that is making the citation.
ii. **ToID**: The ID of the paper that is being cited.

Each row represents a citation record showing that the paper with **FromID** has cited the paper with **ToID**.

This dataset can be useful for analyzing citation networks and understanding the relationships between research papers in High Energy Physics.

Task Description

As a group, you will describe and analyze the citation network in the provided data by addressing the problems below. Your analysis must be based solely on the data given to you. Do not introduce other data into your work. Also, you do not need to investigate the source of the data further.

- Your group will prepare and submit a single, short, structured report that addresses each of the problems set out below.
- All the summary statistics in the report should come from an R program, or be readable from plots in the report.
- Your report and program will be marked by the lecturer, and you may be required to discuss them with the relevant lecturer.
- You will receive group-specific feedback on your submission.

To complete this assignment successfully, you should start your work soon and plan your time carefully.

Problems

Question 1 `[40 marks]`

The **in-citation number** of a paper *i* is the total number of papers *j* such that *j* cites *i*. The **out-citation number** of a paper *i* is the total number of papers *j* such that *i* cites *j*. Using techniques such as summary statistics and plots, describe the in-citation and out-citation numbers of all papers in the dataset. Your description should include both univariate and multivariate analysis.

Question 2 `[30 marks]`

A physicist claims that the length of the bibliography (measured in terms of the number of out-citations) for high-energy physics papers is increasing over the years.

- Compute the average out-citation number in each year from **1993 to 2001**.
- Build an appropriate simple linear model to investigate the physicist's claim.
- Explain your findings in non-technical terms.
- For the linear model, you may assume that the residual plots raise no issues about model assumptions or fit and you should not attempt to analyze or study them (I know that this is not what normally happens, but I am trying to make your life easier).

Question 3 `[30 marks]`

Another physicist claims that people write better papers in the autumn and winter months (**October to March**) than in spring and summer months (**April to September**). Assume that the quality of a paper can be measured in terms of its **in-citation number**.

- Carry out an appropriate statistical test for this physicist's claim.
- Explain your findings in non-technical terms.

Assessment Criteria: how your ICA will be marked

Your exam will be assessed based on the following criteria, with the percentages applied to each question accordingly.

Presentation (30%)

- Clearly and concisely present your work in a well-structured report.
- Maintain a professional and organized style throughout.

Code quality and formatting (50%)

- Write correct, efficient code that meets the assignment requirements.
- Use consistent formatting, including proper indentation, meaningful variable names, and logical organization, to enhance readability.
- Ensure that the code runs smoothly and accounts for potential errors.

Code commenting (20%)

- Provide clear, informative comments explaining the functionality and logic of your code.
- Demonstrate an understanding of how each code segment contributes to the overall solution.
- Keep comments concise yet detailed enough for easy comprehension.

Guidelines for report submission

Name your report `group*.pdf`, where `*` should be replaced with your group number (e.g., for Group 01, it should be `group01.pdf`). The report should be consistent with the following:

- You must use the R markdown template provided on the Moodle page for your report and are not allowed to change its font, font sizes or margins. All the R code used should be visible in the report. If the template has been changed, up to 4% of marks can be lost. Your document will be re-formatted to the template standard, to which the following point will apply.
- The report must not be longer than 5 pages (5 sides) in A4 paper, including figures and R code. Only the first 5 pages of any report will be marked. Note that this doesn't mean that you should aim to fill all the space available to you. Writing more text doesn't necessarily get you more marks. You should keep all of your R code on the PDF report but choose wisely which graphs and R outputs you want to show in the report.
- Please save your report as a PDF file using the R markdown template provided.
- It must be written in clear comprehensible English with readable and well-labelled figures.
- Your report should be anonymous — i.e., there should be no mention of group members' names anywhere in your submission. Use your candidate numbers instead.

Guidelines for Rmd program submission

- Name your R program `group*.rmd`, where `*` should be replaced with your group number (e.g., for Group 01, it should be `group01.rmd`).
- The R code used should be visible in the PDF report as instructed above.
- Assume that the working directory has already been set to the location of the data file and to where any output files will be stored, i.e., there should be no `setwd()` command or reference to directories.
- Import the data from the provided data files.
- Your program will be run using the `knitr()` functions to generate your PDF report. **It should produce the same PDF report submitted, otherwise only the knitted version will be marked and overlength penalties will apply.**
- All parts of the submission should be anonymous — i.e., there should be no mention of group members' names anywhere in your submission. Use your candidate numbers instead.
- Your program should not use non-standard packages (no `library()` command). Only basic R should be used.