SCIF30006 – Advanced Data Science for Scientific Computing

Assessed Exercise – Assessment of Topic 1: Scientific Data Analysis

This exercise will count for 40% of the marks for this unit. It is intended to assess your technical skills when assessing a scientific data set and building simple regression models, and to give you an opportunity to communicate your results in a short report. This relates to the first three ILOs, with a skew towards ILOs 1 and 2, while preparing you for ILO3:

1. Explain the basic steps involved in preparing and curating data and assess data using standard statistical descriptors.
2. Explain different techniques for extracting information from data and select suitable regression models.
3. Describe the basic principles of machine learning, including choice of models and tuning of parameters.

Instructions

In this exercise, we ask you to showcase the following skills:

a) In the first part, **simple linear regression models** are used to derive key scientific data. With models in hand, you will be asked to assess model fit and model predictions with suitable measures and visualisations.

b) In the second part, you compare the model from part one with a more realistic model, **comparing model fit**. You might also wish to consider **data variability**. Your analysis should take account of the scientific context and any problems you identify with the dataset.

c) Along with submitting your code, we also ask you to prepare a **short data analysis report**. In this report, you should justify the chosen methods and models, as well as answering all questions posed. You should highlight key findings and insights, illustrate these with appropriate figures and discuss any limitations and challenges. Appendix 1 provides further details.

You can choose whether to work on the infrared model analysis or the analysis of springs and nanocoils. While a low level of familiarity with the underlying concepts may be comforting, you should not need much domain-specific knowledge. Please select **one** of the topics as early as possible and focus on that, you will only need to submit one option for this assessment.

Submission

Please submit your data analysis report, along with all information needed to check and reproduce your work, to the submission point on Blackboard. This would normally include your report, as well as a separate document of working code which can be run in a Jupyter notebook or a Python IDE, exhibiting best practice, but you could choose to use R or indeed Excel for all or some of your analyses. Provided all work is fully explained and can be repeated, you can choose the platform you find most appropriate.

Make sure that all questions are answered in the data analysis report, that you showcase your skills, and that you fully justify the choices of methods and models made. Marks will be awarded for a clear and concise report, the selection of approaches, clarity of analysis, quality of presentation and clear links with the scientific content, as well as the quality of code and analysis. The marking criteria are broadly similar to those you have encountered in previous years and these are shown in the Marking Scheme document. Code and report components will be weighted equally. We suggest that you spend no more than 40 hours (and ideally considerably less than that) on this assessment. Note that any attempts at "going beyond the brief" need to be full justified in the scientific context and that you need to meet the brief first and foremost.

Clearly indicate which option you have chosen and submit your work by the deadline of Wednesday at noon in week 9 (19th November 2025) on Blackboard. You can find information about extensions and exceptional circumstances on the assessment page for this unit.

If you have any questions about this assessment, please post these to the discussion board. Dr Cameron Beevers will be holding a drop-in session during consolidation week (week 6, 28th October 2025, from 3 pm) and both Cameron and Natalie are also happy for you to contact them by email to request a chat.

AI Tools

Please make sure that you are familiar with current UoB guidance on AI tools; a good starting point is this page: https://www.bristol.ac.uk/students/support/academic-advice/academic-integrity/, and we strongly recommend that you complete the Using AI at University training course. We have summarised the considerations and concerns about AI usage during your studies in Appendix 2.

Formally, this assessment is classified as:

Category 3: Selective. Students may use AI apps (free versions from UK-based app stores only), plug-ins available through the university, apps provided by disability services in line with a bespoke study-support plan, or free-version of web-based generative AI chatbots such as Chat-GPT for the following purposes :

• Provide general feedback on draft work:

"General" here refers to spelling, punctuation, and grammar changes, or enhancements to conciseness and structure. This includes the spelling and grammar checkers embedded into Word or the free version of the Grammarly plug-in. However, you should not allow AI to make changes to the work directly and you should check suggestions that you are unsure about with academic teaching staff.

To get AI to make suggestions rather than perform a full rewrite of your writing use the following in your prompt "Do not rewrite the piece of writing that I am going to give you. I want you to provide both a list of suggestions for improvement and some reflective follow-up questions that I can work through as I go back and edit this piece of work." Note that this is the use of AI to help you edit your own work rather than allowing AI to conduct substantial re-writes on your behalf.

• Preparing code:

When writing computer code, you may sometimes find yourself entering problematic sections into a search engine to help you with debugging it. It is then possible that you will get suggestions from discussion boards, such as Stack Overflow, but there may also be AI generated sections. Similarly, when using AI as a tutor, it may offer help with code segments. You should not be asking AI to generate code for you directly but rather suggest possible edits or corrections to make it work properly. We expect you to check these suggestions, making sure that you understand the code, and to cite such sources appropriately.

To get AI to identify potential bugs rather than rewrite your code, you could use the following in your prompt:

Please review the following code:

[paste code here]

Consider potential bugs. Please suggest improvements and explain your reasoning for each suggestion.

You are ultimately responsible for making sure that you can properly explain code in terms of how it works, its performance on selected inputs, and why you structured your code in a particular way.

• Translation software:

Part of studying a degree in the UK means that students are expected to produce written work in English. The written assessments are designed to assess how clearly a student puts together their thoughts to showcase their scientific understanding of both chemical theory and their results. Minor errors are a common occurrence that affect everyone, especially in early drafts of a written piece of work. However, in some cases, especially early on in the degree, some students might benefit from composing their thoughts in their native tongue then translating their work into English. The free version of apps that make suggestions on acceptable word substitutions and/or provide grammatical suggestions may be used in such cases. However, you should not allow AI to make changes to the work directly and you should check suggestions that you are unsure about with academic teaching staff.

Students are responsible for the accuracy of any translations of their work. All students who are not native English speakers should also be utilising university resources to build up their technical English proficiency so that they can move away from direct translation to writing directly in English before the end of their degree. Your personal tutor and/or the unit directors can help you identify which university resources would be most helpful so please make sure to reach out to them for advice.

• Reference management:

You can use a reference manager such as EndNote, Mendeley, or Zotero to organise your references, insert them into your work, and generate a reference section at the end of your dissertation.

You are still responsible for the accuracy of the reference citations and you should check the generated references and correct any that are incorrect or incomplete.

• Citation of Technology Used:

Acknowledgement of the type and version of all computer applications (e.g. Word, Excel, PowerPoint, Jupyter, LaTeX, etc) and AI (e.g., ChatGPT, Gemini, Claude, co-pilot, built-in tools for Word, etc.) used in drafting, editing, or correcting of any material submitted for assessment. Two example declarations are given below:

This document was prepared in Microsoft Word for Microsoft 365 using the included spelling and grammar checkers. The data were analysed and graphed using Microsoft Excel for Microsoft 365. The linear regression analyses of plotted data were completed using the "Analysis Toolpak" add-on for Microsoft Excel for Microsoft 365. The writing was edited using suggestions from both the free-version of Grammarly and Chat-GPT4.

This document was prepared in Microsoft Word for Microsoft 365 using the included spelling and grammar checkers. The datasets were analysed in Jupyter Notebook 7.4 from Anaconda using Python 3.13. The python code produced was debugged using suggestions from Stack Overflow (reference relevant discussion threads) and Chat-GPT4 in line with the teaching labs AI policy document. The writing was edited using the free-version of Grammarly.

Unspecified Usages of AI

Outside of the above, the use of generative AI/large language models including Chat-GPT, Bard, Gemini, CoPilot, Dall-E, PI, and Llama2 is prohibited on this unit.

Any work suspected of being generated either in part or in whole through the unauthorised use of AI or through an authorised use of AI but without appropriate acknowledgement, which would be considered making false claims, will be referred for further investigation of academic misconduct in line with university policy.

Data and Context

Please choose ONE of the options for your work.

Option 1: Infrared Models

Infrared (IR) spectra are often analysed through application of a simple harmonic oscillator model, which conforms to the equation $F=-kx$ where k is the force constant, F is the force, and x is the displacement from equilibrium. The force constant, k, gives a measure of the stiffness of the bond, which relates to the frequency of the infrared mode.

a) A number of bond stretching modes have been modelled to examine the forces acting on bonds which have different force constants. Force constants have been derived from measured IR frequencies. The file Option1_part_a.csv contains data for displacements, x, and forces, F. Clean the dataset to remove missing data and extreme outliers. Generate appropriate plots to examine to what extent the data for different types of bonds are described well by the harmonic oscillator model. The force constant is a measure of bond stiffness and can be determined through application of simple linear regression, where the slope of a plot gives k values in units of nano Newton per Ångstrom. Determine the force constants for the different bonds included in this dataset and evaluate the model fit. Do the prediction errors (residuals) of your linear model appear biased relative to the dataset values?

b) The file Option1_part_b.csv is a dataset of IR bond stretching mode data that have been generated from a model based on a Morse potential, which is more complex to evaluate but can allow for bond-breaking.[1] The Morse potential is generally considered as more realistic than a harmonic oscillator, and here we would like you to assess whether this is the case.

Strain is a unitless quantity calculated as strain = |x|/equilibrium length. Strain = 0 is equivalent to the equilibrium bond length. Apply a linear fit to the medium to high strain regime, where strain <= 0.5, and a low strain regime, where strain <= 0.05. Use these linear models to determine k values for the different strain subsets, and evaluate each model. Analyse the fit of these models for the bonds in each subset and discuss whether the prediction errors appear biased relative to the Morse potential values provided in the dataset. The displacements of IR vibrations are usually considered to be between 1 % and 3 % of equilibrium bond length. By referring to the above analysis, consider whether the simple harmonic oscillator model is suitable for analysing infrared stretching modes in this dataset.

References

1. *Atkins' Physical Chemistry*, Oxford University Press, Oxford, Twelfth edition., 2023, pp. 452–455.

Option 2: Hookean Springs and Nanocoils

Hookean springs conform to Hooke's law, where the force restoring a spring to equilibrium is given by *F*=-*kx* (*F* is the restorative force, *k* is the spring constant, and *x* is the displacement from equilibrium). Some nanocoils have been found to act as Hookean springs within a low strain limit.[2] Strain = |x|/equilibrium length, where strain = 0 is equivalent to the spring length at equilibrium.

a) A number of springs have been modelled to examine the forces that act upon springs with different force constants. The file Option2_part_a.csv contains data for the displacements and forces of these model springs. Clean the dataset to remove missing data and extreme outliers. Generate appropriate plots to examine to what extent the data for different types of springs are described well by the harmonic oscillator model. The force constant is a measure of spring stiffness and can be determined through application of simple linear regression, where the slope of a plot gives *k* values in units of nano Newton per Ångstrom. Determine the force constants for the springs included in this dataset and evaluate the model fit. Do the prediction errors of your linear model appear biased relative to the dataset values?

b) The file Option2_part_b.csv is a dataset of nanocoil stretching that has been generated from a model using a non-linear approximation of spring behaviour. The potential used in this model is generally considered more realistic than a harmonic oscillator, and here we would like you to assess whether this is the case.

Strain is a unitless quantity calculated as strain = |x|/equilibrium length. Strain = 0 is equivalent to the equilibrium length. Apply a linear fit to the medium to high strain regime, where strain <= 0.1, and to the low strain regime, where strain <= 0.01. Use these linear models to determine *k* values for the different strain subsets, and evaluate the quality of fit for each model. Analyse the fit of these models for each strain subset and discuss whether the prediction errors in the model appear biased relative to the potential values provided in the dataset.

The low-strain limit is usually considered to be less than 1 % of the equilibrium length. By referring to the above analysis and assuming real nanocoil springs conform to the second dataset, consider whether the simple harmonic oscillator model is suitable for modelling nanocoil springs within the low-strain limit.

References

2. H. Zhan, G. Zhang, C. Yang and Y. Gu, Breakdown of Hooke's law at the nanoscale – 2D material-based nanosprings, *Nanoscale*, 2018, 10, 18961–18968.

Appendix 1 Data Analysis Report

For this exercise, we ask you to write a short data analysis report about your work. This is designed to encourage you to justify which approaches you have chosen to analyse your data and to assess the models fitted. Note that we do not want you to include every possible option in the code submitted, but to be selective – the report allows you to explain your choices. In addition, please make sure you provide answers and appropriate illustrations for your key findings in this report. Unlike a full project report, you do not need to provide an extensive introduction to the topic, but your report may refer to key literature references used for comparison.

Please use the following structure for your report:

1. Methodology – this includes any data pre-processing that you do, as well as giving an overview of the methods and models you have chosen. There is no need to include code examples, but aim to give enough information to allow somebody else to reproduce your work with a good understanding of what you did.
2. Results and Discussion – while you may choose to split this into 2 sections (first showing all of your results and then a discussion section), it can be easier to combine these into a single section; this may help you to address the different parts of the exercise, too. You should present key findings and insights, making sure that these address all questions asked and develop a clear story. This section should be illustrated with suitable figures and your results should be critically evaluated, highlighting any limitations and challenges.
3. Conclusion – use this to give a concise summary of key points from your work and highlight the insights developed.

The code submitted forms part of your assessment and you do not need to reproduce it in your report. It should be functional and annotated with comments and/or markdown boxes so markers can follow your process.

Please aim to provide a clear and concise report – with figures, it becomes difficult to be prescriptive about actual length, but we suggest that you avoid going much beyond 2 sides of A4 in terms of the written text. Since this is a new report format for you, this will not be enforced. We are looking for a correct and thorough analysis, showing your understanding of both the underlying scientific problem and the data analysis techniques which you use. The report should be well-written, rather than a list of bullet points, it should focus on the data analysis and questions, and where relevant, it should include appropriate references. There is no need for an abstract or a lengthy introduction, you can use the 3 sections specified above.

Appendix 2: Further Information on AI Usage

Since 2022, the rise of generative AI chatbots such as Chat-GPT, CoPilot, Bard, Dall-E, PI, Claude, Llama2 means that scientists can utilise AI in new ways to enhance their outputs. However, the use of AI to produce even a short response (100 words) requires significant amounts of both electrical power and fresh water as summarised in the article Chat-GPT Energy Consumption Visualised. Unfortunately, scientists and students can also utilise AI in a manner that allows them to manipulate their data to provide a skewed picture of their results or produce written pieces of work, either as a whole or in part, that can be falsely attributed to a human author.

All of the assessments in the practical units are designed so that they can be completed by students to a high standard without the use of generative AI. There are some problems related to using AI freely to complete your assessed work:

•Using Chat-GPT and related programs to wholly prepare your assessments means that the work of that generative AI will be evaluated as if it were your own work. This is considered contract cheating by the University of Bristol and will be subjected to referral for academic misconduct in line with the university policy on academic misconduct.

• Significant AI usage can result in brain changes that impact your ability to engage in "neural intensive" tasks. There is a study from MIT that explored changes in neural mapping in heavy users of AI, and they saw significant neural pruning and pathways collapse that affected users ability to recall their outputs and think critically.

•Using AI for "lower-level" tasks such as summarising journal articles or textbook chapters means that you are not fostering the skills to rapidly condense a technical paper or document into a concise executive summary that can be incorporated in your own writing. Additionally, quickly skimming an AI-generated summary instead of reading the paper yourself means that you will miss out on important details contained in the full document.

•In terms of behaving with integrity, note that any information included in a session with most AI models becomes part of the training data used to refine current and future AI models. This has significant implications if you upload copyrighted materials or materials with a limited acceptable use, such as the bespoke academic materials you utilise in your degree courses.

•For the physical sciences, AI is prone to including hallucinations and wrong information that can results in users, who do not fact check with textbooks or published journal articles, internalising misconceptions that can have a significant negative impact on their learning.

•Relying fully on AI to complete your assessments also means that you are stunting your growth as a scientist and you will have misrepresented your ability on key transferrable skills when you apply for jobs and/or further academic study.

However, in limited instances AI can be a useful tool to help you refine your work and foster your skill development. For more information about acceptable and unacceptable uses of AI you should complete the Using AI at University training course prepared by the University of Bristol.