From research question to model specification:

Smart modelling strategies and critical use of generative AI

Dates: 11th, 12th & 15th June, 09:00 to 17:00

Location: Life and Mind Building, South Parks Road, Oxford, OX1 3EL

Course description

This course is designed for D.Phil students who are planning a quantitative analysis and need support in moving from their research questions and data structure to a well-specified statistical model.

A common problem in doctoral work is that students may know the broad method they intend to use, but they are much less confident about how to specify the model in practice. How should continuous predictors be represented? When should non-linearity be considered? Which interactions are scientifically meaningful and realistically estimable? What is the appropriate level of complexity given the amount and structure of the data? What assumptions are being made, and are they compatible with the dataset?

The course addresses these questions through a structured workflow that starts from the substantive research aim, moves to the statistical model, and then returns to interpretation of the answer. A second key component is the critical use of generative AI tools. Students are increasingly using tools such as ChatGPT or other GenAI systems for coding and analysis support, but these tools are only useful when the user can provide a clear modelling objective, sensible constraints, and enough contextual information. Generative AI can generate many plausible options, but it does not reliably recognise the scientific priorities, data limitations, or modelling constraints of a given D.Phil project. For that reason, students need to decide first what kind of model they are trying to build, what form of output they need, and what modelling choices are feasible given their data.

The course therefore combines:

1. core principles of smart model specification
2. a framework for deciding how to parameterise a model
3. guided use of generative AI for prompting, iterating, and critiquing analysis suggestions
4. discussion of when AI-generated modelling advice is sensible, superficial, or simply wrong

The emphasis is practical. Ideally, students will work with their own data and analysis plans. Where this is not possible, structured example datasets will be provided.

Target students

The course is aimed at D. Phil students at any stage of their studies.

How this training supports D.Phil work

This training supports D.Phil students by helping them:

- translate a substantive research question into an analysable statistical model
- make defensible modelling decisions rather than relying on default settings or convenience choices
- understand the consequences of alternative parameterisations for interpretation
- use generative AI tools more effectively by supplying better prompts, clearer goals, and stronger constraints
- critically evaluate AI-generated code, modelling suggestions, and interpretations rather than accepting them at face value
- improve the quality, transparency, and efficiency of their quantitative analysis

Learning aims

By the end of the course, students should be able to:

1. explain what model specification / parameterisation is and why it matters
2. move systematically from research question and study design to a candidate model structure
3. make principled choices about how predictors should enter a model, including non-linearity, interactions, transformations, grouping, and simplification
4. recognise when a modelling idea is not compatible with the available data
5. use generative AI tools to support analysis planning and coding in a more targeted and efficient way
6. critically assess AI-generated modelling suggestions and identify weak or misleading recommendations
7. articulate and justify modelling choices for their own D.Phil analysis

Skills gap addressed

Many doctoral students are taught statistical methods as named techniques, but not how to make the modelling decisions that determine whether an analysis is actually useful. They may know that a given outcome suggests a logistic, Poisson, or mixed-effects model, but they are rarely taught how to decide:

- how continuous variables should be represented
- whether interactions are warranted
- when flexibility is helpful and when it is overfitting
- how to balance scientific meaning, statistical identifiability, and data limitations
- how to communicate these choices clearly

This skills gap is now even more important because students increasingly use generative AI for coding and analysis support. Without a clear prior understanding of the modelling problem, students are easily overwhelmed by the large number of plausible AI-generated options. The course addresses both gaps together: model specification itself, and the critical use of GenAI in that process.

Transferable skills

The course develops several transferable skills relevant beyond the immediate D.Phil project:

- structured analytical thinking
- formalising research questions into decision steps
- critical evaluation of automated outputs
- prompt design for technical tasks
- documentation and justification of modelling choices
- communication of quantitative reasoning to non-specialists and collaborators
- reproducible and transparent research practice

For further information:

Dr. Matteo Tanadini and Dr. Claude Renaux

info@zurich-data-scientists.ch

Zurich Data Scientists GmbH