STAT2120: Linear models 2025/26
Project

Important:

- Work in groups of two or three people.
- Submit a written report by **Monday, January 5, 2026, 12h**, recto-verso, by email (pdf only) to the assistant and the professor. Attach also your R or python codes to the email in separate files.
- Validate the dataset and the composition of your group by email to the assistant before **Friday, November 14, 2025**.
- The title page of the report contains the names of the group members, a title of your project that describes the objective, the course number and the academic year.
- Put graphics and tables either in the text, or appendix.
- Page limit: 20 pages altogether. Text: maximum 15 pages, Appendix: maximum 5 pages.

Data:

Every group uses its own dataset, to be validated by the assistant. Try to find a meaningful empirical study, where you determine a response variable Y and the explanatory variables X. You must have at least seven quantitative variables, and at least two qualitative variables.

Tasks:

1. Separate arbitrarily a small subset of observations from the dataset (e.g. 10 %). These observations will be used for validation and prediction, but not for model estimation.
2. Start your report by formulating an objective. What is the idea and the goal of your project? At the end of the report, explain why or why not the objective has been attained.

3. Do a descriptive analysis of the variables of the model. Provide a table with mean, standard deviation, skewness and kurtosis. Show boxplots of the variables, and the correlation matrix.

4. Select an adequate model for the response variable by considering all quantitative and qualitative variables and by using our model selection strategies. Consider also possible interactions of the qualitative variables with one or several quantitative variables. Verify the underlying hypotheses and, if necessary, take remedial actions. For example, check for

   (a) nonlinearity
   (b) outliers and/or influential observations,
   (c) multicollinearity,
   (d) heteroskedasticty, and
   (e) autocorrelation.

   If necessary, try to improve the model by using the methods seen in class. If different models are chosen for different methods, select the one that minimizes the mean squared prediction error using the validation data.

5. Test for significance of the estimated coefficients of the obtained model, and interpret their signs. Give a more detailed interpretation of the coefficients of the qualitative variables.

6. Test a linear combination of at least two coefficients, which makes sense in the context of your project.

7. Test a subset of coefficients (at least two), for example corresponding to the qualitative variables, to be equal to zero.

8. Calculate prediction intervals for the validation data. Does the coverage percentage of these intervals correspond to the nominal level? If not, what could be a reason?

9. Give a general conclusion about the adequateness and usefulness of your model.

10. Indicate clearly for which parts of the project, to what extent and for which purpose you have used artificial intelligence such as LLM for your work, see the following guidelines.

Use of Artificial Intelligence:

Various types of use are permitted, some of which do not need to be supported by a reference, while others do. When the use of an AI tool requires a reference, students are encouraged to describe in an appendix (to their assignment, report, etc.) how the AI was used.

1. Examples of uses that do not require any indication that an AI tool was used

   (a) Use of a language assistant tool. Students may use language assistance functions without any mention. This includes both dedicated tools such as Grammarly, Antidote or DeepL and generic tools such as ChatGPT. If the source text has been written by the student, an AI tool may be used to correct spelling, grammar and syntax, to search for synonyms, etc. A tool such as DeepL may be used to produce a first version of a translation (which is subsequently revised), for example of a text that the student has written in another language; however, the tool must not add any content not written by the student.

   (b) Use of a search engine or brainstorming tool. Students may use AI tools without citation for exploratory purposes, to find their way into a subject or to identify lines of enquiry worth pursuing or relevant sources (facilitated by the integration of search tools into generative AI). This form of information-gathering is similar to the use of a search engine (Google, Bing, DuckDuckGo or Qwant, etc.). Such uses are permitted as long as the student checks the comments and references that are encountered, looks up the sources identified and analyses them. Although the use of the tool does not need to be indicated in such cases, students must of course provide proper references for the consulted sources to which the tool has given access. More generally, it should be borne in mind that any cited source must have been checked. It is therefore unacceptable to copy information provided by an AI tool, whether in the text or in a bibliography, for example. It may in fact be risky to do so, as these tools often invent sources. It is also unacceptable to confine oneself to sources identified using an AI tool. If the requested assignment requires bibliographical research, relevant databases and library catalogues must be

searched in accordance with standard practice.

(c) Help with code documentation, e.g. the specification of methods, unless this is explicitly prohibited.

2. Examples of uses that require indication of which AI tool was used

(a) Citations. If a student uses verbatim content generated by an AI tool (not only text, but also illustrations produced by an image generator), the tool should be cited, for example in a specific section or a footnote; a distinction should be made with bibliographical references in such cases. Content generated by an AI tool with a source citation must be limited in comparison with the overall work (just as quoted text must remain proportionate to the size of an academic work).

(b) Source code. Any fragment of code that has been used and is submitted in work containing a significant element of code (e.g., a method or the implementation of an algorithm) must be introduced with an indication that the fragment has been generated by AI. This is the same rule that applies when code has been copied and pasted from another source code with a compatible licence.

(c) Translations. Students using a text that has been automatically translated from a source written in another language must indicate the source used (in its original language version) in accordance with the citation guide and also indicate: 'translated into [language, e.g. French] with [the AI tool, e.g. ChatGPT (used on 20 November 2023)]', unless they have substantially modified the translation (in which case the tool has merely been used as a language assistant).