Bayesian Statistics 105.173

Data analysis project Winter 2023

Project details

- The final project is to be presented on **January 24th, 2024**.
- Students should upload the assignment on TUWEL by **23:59 on January 23th, 2024**.
- The project can be done in groups of **1-3 students**. The best group size is 3: you learn more from discussions with others; team work is also key in industry.
- Groups of 3 students are expected to do more work than groups of 1 or 2 students.

Project report

Each group will write and submit a report. Guidelines for writing the report:

- Reports should clearly present the **problem and results** of the data analyses.
- Reports should not only include code and figures.
- The report should also include information about the workflow steps (aka, models used and the rationale behind them, hyper-parameters used in the prior, model fitting, prediction, model checking, model comparison, etc).
- Diagnostic outputs and code should be included in the appendix.
- The main report (including figures, but excluding the appendix) should ideally be around 10 pages and, in any case, not exceed 20 pages.
- It is not allowed for different groups to work together or cheat.
- The use of AI is permitted. However, the group members should clearly state in which steps AI was used.

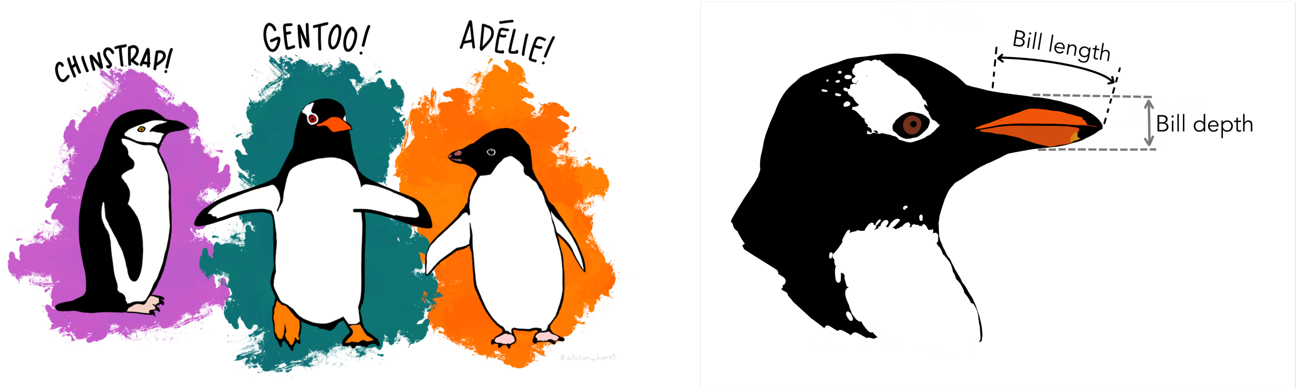
Presentation

In addition to submitting report, each group will present their work in class. Guidelines for the presentation:

- The duration of the presentation will be **10 minutes** (groups of 1-2 students) or **15 minutes** (groups of 3 students), followed by 5-10 minutes of questions.
- The presentation should be quite general. However, students should be ready to reply to specific questions about the project (it is then advisable to have extra slides at the end).
- Bonus points will be given to creative presentations.
- The presentations should contain:
  - Group members' names
  - Data visualization and data exploration
  - Description of at least two possible models
  - The chosen statistical model and justification of priors, hyper-parameters, etc
  - Plots with easy-to-read labels
  - A summary on the last slide

Antarctic Penguins

The goal of this assignment is to analyze the data collected by Gorman et al. (2014), focusing on three types of penguins: Chinstrap, Gentoo and Adélie, displayed in Figure 1a. We want to investigate if we can create a model to predict the bill length of these penguins (see Figure 1b). The bill length or culmen is one of the most important morphological indicators of Antartic penguins (Lee et al., 2015).



(a) Antartic penguins species

(b) Bill morphology

Figure 1: Antartic penguins illustrations. Artwork by @allison_horst

Data can be downloaded from TUWEL and load in R as:

```
> penguins = readRDS("penguins.RDS")
> head(penguins)
  species    sex bill_depth bill_length
1  Adelie   male       18.7        39.1
2  Adelie female       17.4        39.5
3  Adelie female       18.0        40.3
5  Adelie female       19.3        36.7
6  Adelie   male       20.6        39.3
7  Adelie female       17.8        38.9
```

(1) **Data exploration**: Provide a visual description of the data, explaining what you see and what you expect to find out.

(2) **Modelling**:

  (a) Give a mathematical description of at least two linear models, of which at least one should be non-hierarchical (separate or pooling) and at least one hierarchical.

  (b) Explain the different plausible observation models and the variable selection criteria you used.

  (c) Every parameter needs to have an explicit proper prior. **Improper flat priors are not allowed**.

  (d) Explain if you used informative or weakly informative priors, and provide a justification of the choice of these priors and the corresponding hyper-parameters.

  (e) The model should be fit on Stan; `brms` and `rstanarm` R-packages are not allowed.

(3) **Model checking**: Provide posterior predictive checks, and evaluate your models. This should be reported for all models.

(4) **Model comparison**: Compare your models (e.g. with LOO-CV).

(5) **Sensitivity analysis**: Perform a sensitivity analysis with respect to prior choices (i.e. checking whether the result changes significantly when changing the prior). This should be reported for all models.

(6) **Discussion** of issues and potential improvements.

(7) **Conclusion**.

(8) **AI disclosure**: If you used AI, clearly state where and why you used it.

(9) **Appendix** with the Stan `R`-code and outputs; if necessary, additional figures. If you upload your code on github you get bonus points.

References

Gorman, K. B., Williams, T. D., and Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of antarctic penguins (genus pygoscelis). *PloS one*, 9(3):e90081–e90081.

Lee, W. Y., Jung, J.-W., Han, Y.-D., Chung, H., and Kim, J.-H. (2015). A new sex determination method using morphological traits in adult chinstrap and gentoo penguins on king george island, antarctica. *Animal cells and systems*, 19(2):156–159.