

# CS-E407520 - Special Course in Machine Learning and Data Science: Bayesian Workflows

Session 5: Model Expansion and Selection

---

May 20, 2024



## Schedule for today's session

---

Time	Activity
15 min	Peer discussion
30 min	Discussion of workflow diaries
10 min	Break
15 min	Primer for next workflow steps
20 min	Question session

---

# Questions!

At any point during the session, you can submit a question for discussion! Simply go to:

- <https://presemo.aalto.fi/workflowcourse2024>

## Peer Discussions

Form groups of two or more and discuss:

- Your progress during the week
- Any issues you encountered (and how you resolved them, if relevant)
- Questions or feedback you have for the session

## Discussion of workflow diaries

When it is your turn, please

1. present your workflow diary via projector or screen
2. briefly summarise your data and research question
3. teaching staff will guide discussion of your implementation of the workflow steps

## Your goals were:

- Propose, justify, and fit at least two expanded or alternative models
- Choose an approach for selection (prediction vs explanation)
- Perform model selection using an appropriate method
- Summarise results in a single table

## Let's take a break! (10 min)

Some suggestions for recharging during breaks :

- move your body
- open a window or go outside
- drink some water
- try to avoid checking e-mails, messengers, or social media

# **Primer: Inference and Presentation of Model Quantities**

---



# Primer: Inference and Presentation of Model Quantities

- What is your goal/question?
  - What post-processing is needed from your model to answer this
  - Tailor your methods and presentation accordingly
- Population Inference
  - What does our model imply about the wider population or future data?
- Causal Inference
  - Key goal for many studies
    - Did the treatment cause a recovery?
    - Did the movie 'Sideways' cause a reduction in merlot sales? (Yes)

# Population Inference - Bayesian Poststratification

- Poststratification is a common approach to population inference
  - Many, many, extensions
- Using results and characteristics from current sample to draw inferences about broader population
  - Group sample by combinations of adjustment variables
  - Estimate predicted probability/outcome for each group
  - Use the population counts for each group to calculate a weighted estimate
  - Repeat for every MCMC draw to get both estimate and uncertainty
- Let's have a look at an example: Poststratification

## Causal Inference - Counterfactuals

- Probability of different outcome given different causal effect
  - Or, the probability that the observed outcome was not due to the causal effect
- Using the posterior from the fitted model, change the causal effect
  - Gives predicted probability and uncertainty

# Presenting Results

- Effectively presenting model results requires considering both research goal and target audience
  - Research goal: What quantities do you need to present?
  - Audience: How can you convey what the quantities “mean” and justify your conclusions?

## Presenting Results - Goals/Quantities

- Less is more!
  - Modelling process can involve many quantities/processes/steps
  - Not all of these are relevant (or interesting) for all audiences
- Questions to consider:
  - What part of your results answers your question?
  - What is the level of uncertainty in your results?
  - Did you have to make any significant adjustments to your model?

## Presenting Results - Audience

- Different audiences will “need” different presentations of results
- Academic Journal
  - Supporting statistics are essential - (convergence, ESS, credibility intervals)
  - Plotting (generally) static and grayscale
  - Presentation should be structured
- Informal
  - More emphasis on visual appeal/accessibility over technical detail
  - Flexibility in approach to presentation
    - Interactivity with Shiny apps, plotly
    - Ability to embed analysis - jupyter notebooks
- Match formatting/style to audience

## Your Goals for the Next Week

- Choose an “Audience”:
  - Academic journal/writing
  - Informal presentation (e.g., case study)
- Prepare a brief summary of your model’s results, considering your chosen audience
  - No more than 1 paragraph

# Resources for the Week

## Readings

- Workflow Book:
  - Ch.4: Inference for quantities of interest
  - Ch.27: Using a fitted model for decision analysis: Mixture model for time series competition
- Stan User's Guide - Poststratification

## Resources

- Talk: Matthew Kay - A biased tour of the uncertainty visualization zoo
- Digits Case Study

## Relevant Papers

- Rohrer, Schmukle, & McElreath; The only thing that can stop bad causal inference is good causal inference