Introduction(week 1, part 1)

2 Sep 2023

Table of contents

Basics

Logistics

- (almost) everything at course web page, https://bbolker.github.io/stat720
- communication/forums (Piazza)
- assignment marks (Avenue)
- Zoom/recordings (by request)

Integrity

- notes on honesty
- why copying code is good
- Stack Overflow, ChatGPT, and all that
- group work

Prerequisites

From the course outline:

- basics of linear models (as in STATS 3A03), with associated linear algebra
- basics of generalized linear models (as in STATS 4C03/6C03), including knowledge of exponential family distributions
- inferential statistics: sampling distributions, Central Limit theorem, hypothesis testing, Wald tests, maximum likelihood estimation
- ideally, *basic* knowledge of Bayesian statistics and Markov chain Monte Carlo estimation
- intermediate knowledge of R

Goals

- principles/practices of statistical modeling
  - choosing a model
  - diagnostics and troubleshooting
- good intermediate understanding of the tools (ridge/lasso, (G)(LA)MMs; unifying principles
- awareness of computational foundations/scaling

Technical skills & tools

Not focal, but unavoidable and useful

- R (base + some tidyverse)

- reproducibility
  - version control (Git/GitHub)
  - documents: Quarto/Sweave/Jupyter notebooks

about me

- weird background (physics/math u/g, Zoology PhD, epidemiological modeling)
- math biology (ecology/evolution/epidemiology)
- computational statistics (mixed models, Bayesian stats)

things I like/obsess about

- scientific inference $\gg$ pure prediction (but see Navarro (2019))
- generative models
- data visualization
- solving problems in context, practical issues
- bad statistical practice (p-value abuse, snooping, dichotomania, imbalance handling, ...)

Navarro, Danielle. 2019. "Science and Statistics." Aarhus University. https://slides.com/djnavarro/scienceandstatistics.

Modeling loop

cyclic process

but beware the garden of forking paths!

- "researcher degrees of freedom", "HARKing", etc.
- Simmons, Nelson, and Simonsohn (2011); Gelman and Loken (2014)

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science: Data-Dependent Analysis–a "Garden of Forking Paths"–Explains Why Many Statistically Significant Comparisons Don't Hold Up." *American Scientist* 102 (6): 460–60. http://link.galegroup.com/apps/doc/A389260653/AONE?u=ocul_mcmaster&sid=AONE&xid=4f4562c0.

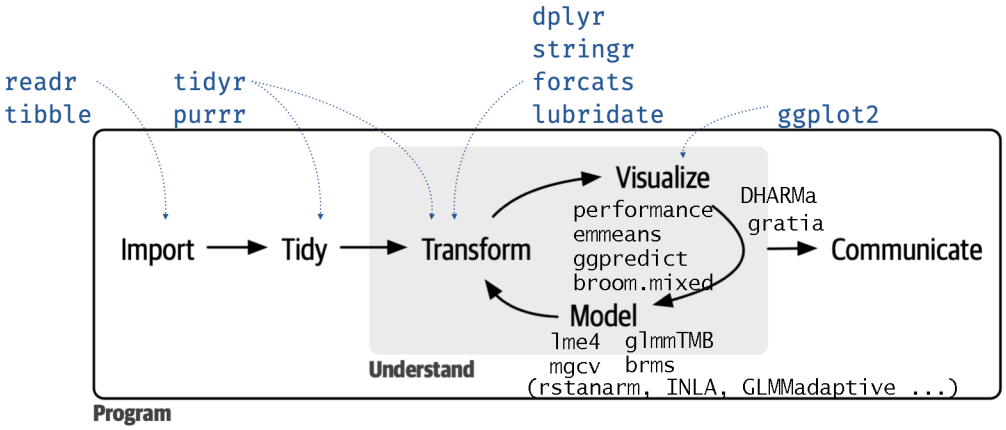

Figure 1: original from Mine Çetinkaya-Rundel

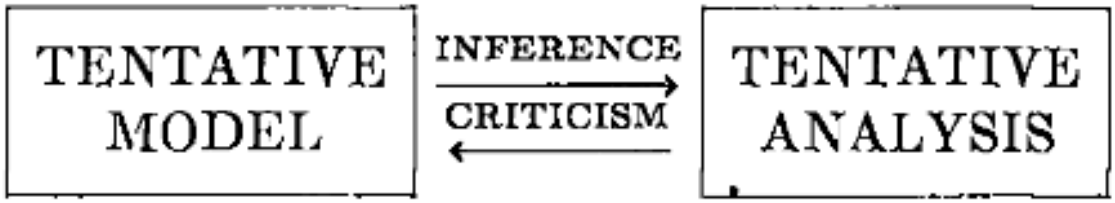

Figure 2: From Box (1976)

Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71 (356): 791–99. https://doi.org/10.1080/01

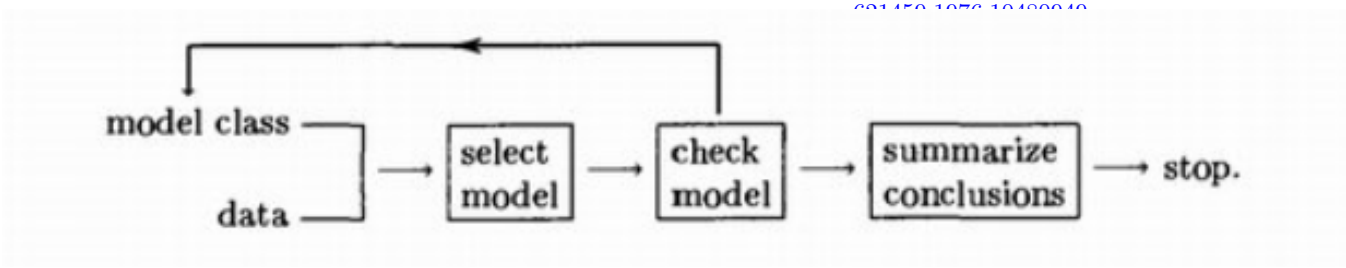

Figure 3: From McCullagh and Nelder (1989) p. 392: 'The introduction of this loop changes profoundly the process of analysis and the reliability of the final models found.'

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. London: Chapman; Hall.

Figure 4: from Art Share LA