Introduction(week 1, part 1)

5 Sep 2024

Basics

Land acknowledgement

McMaster University (and this class) is on the traditional territories of the Mississauga and Haudenosaunee nations, and within the lands protected by the "Dish with One Spoon" wampum agreement. (**Why?** See "Beyond Territorial Acknowledgements")

Logistics

- (almost) everything at the course web page
- communication/forums (Piazza); e-mail if necessary
- assignment marks (Avenue)
- Zoom/recordings (can be set up 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
  - interpreting and communicating results
- understanding the tools ((penalized) (G)L(A)(M)Ms; unifying principles of regression modeling, shrinkage/penalized estimators
  - both frequentist and Bayesian approaches
- awareness of computational foundations/scaling

Scope

- **regression** modeling
- the components:
  - linear model matrices/basis spaces
  - link functions
  - conditional distributions (GLM "families")
  - penalization/shrinkage
- includes a vast range of useful models
- **not** clustering, neural nets, tree-based models

Technical skills & tools

Not focal, but unavoidable and useful

- R (base + some tidyverse)
- reproducibility (Jenny Bryan (2017); Jennifer Bryan (2017))
  - version control (Git/GitHub)
  - documents: Quarto

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))
- data visualization
- solving problems in context, practical issues
- battling bad statistical practice (p-value abuse, snooping, dichotomania, imbalance handling, ...)

a tiny bit of philosophy

- frequentist vs. Bayesian
- *computational* or *pragmatic* Bayesian
- in a perfect world either choice would be available for any problem
  - Bayesian approaches generally more flexible but slower
  - priors!
- hierarchical/latent/mixed models make the boundaries fuzzy (*empirical Bayesian*)

strong feelings about p-values

- please use frequentist methods correctly
- H_0 is almost never true
- **never** accept the null hypothesis (use *equivalence tests* if you really care)
- a low p value doesn't mean the effect is large or important
- may help to frame results in terms of **clarity** (Dushoff, Kain, and Bolker 2018)

The modeling cycle

Before you start

- you need to know what the question is!
- this is hard for statisticians
  - when mining data, go back to the original paper(s)
- a low p-value is not inherently interesting!
- what is a large effect? what is an interesting effect?

From X/Twitter

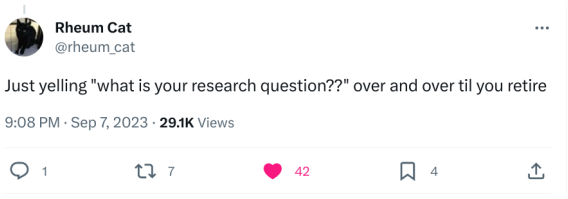

Figure 1: from user rheum_cat on X

Effect sizes

- standardized measures like Cohen's d ($(\bar{x}_1 - \bar{x}_2)/s$, where s is some measure of pooled standard deviation: Wikipedia) are common ...
- but shouldn't be used mindlessly. Real-world, unstandardized effects are usually more meaningful
- effects estimated on the log or logit scale are unitless and hence *may* be easier to generalize
- scaling predictors and responses may help (Schielzeth 2010)

An iterative process ...

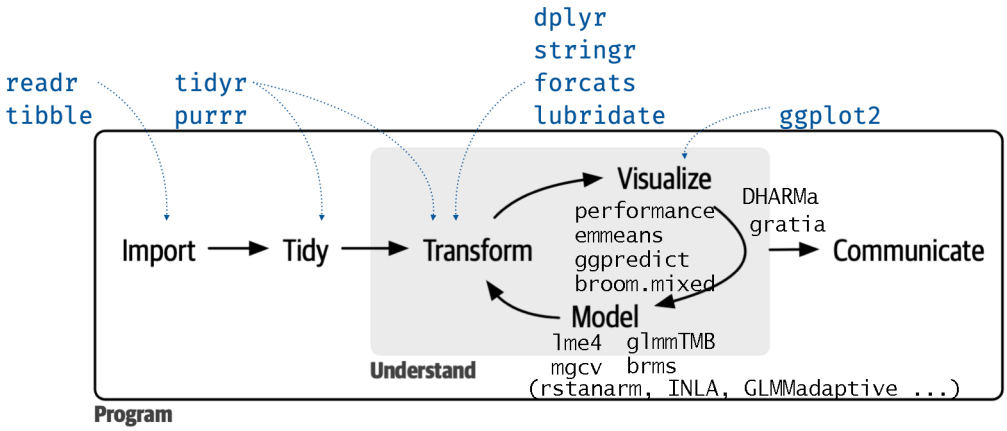

Figure 2: original from Mine Çetinkaya-Rundel

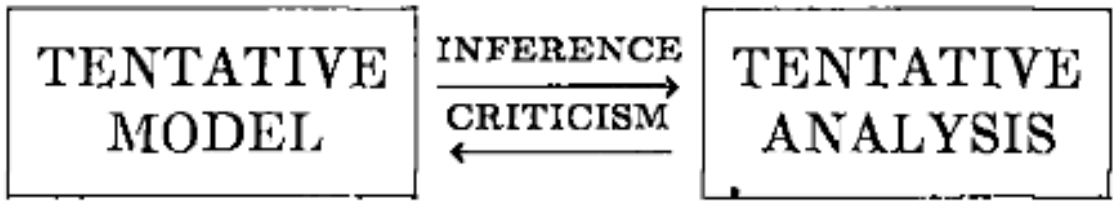

Figure 3: From Box (1976)

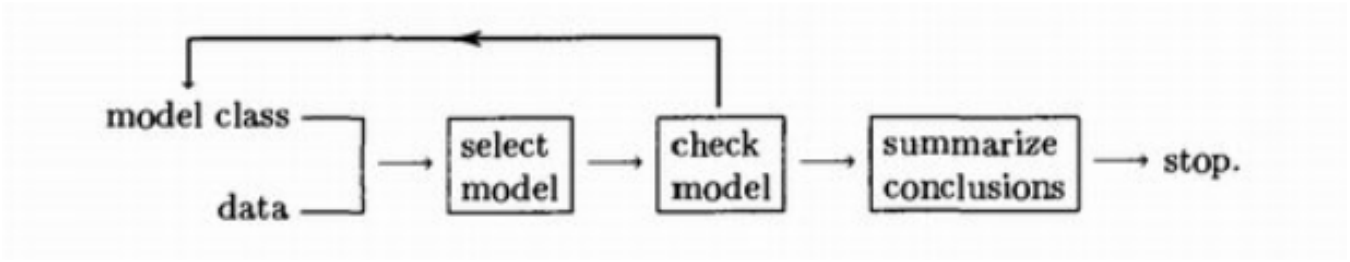

Figure 4: From McCullagh and Nelder (1989) p. 392: 'The introduction of this loop changes profoundly the process of analysis and the reliability of the final models found.'

Beware the garden of forking paths!

Figure 5: from Art Share LA

- "researcher degrees of freedom", "HARKing", etc.
- Simmons, Nelson, and Simonsohn (2011); Gelman and Loken (2014)
- "forking paths" < snooping, HARKing < p-hacking
- compare **data leakage** in predictive modeling (Hastie, Tibshirani, and Friedman (2009) §7.10.2, Kapoor and Narayanan (2023), Bussola et al. (2020), Quinn (2021))

Solutions?

- pre-registration (formal or informal); report deviations from planned analysis
- choose model complexity (see Harrell RMS ch. 4), do diagnostics etc., **without reference to response variable** or metrics of significance
- *initial* data analysis (response-free): Heinze et al. (2024)

Choosing model complexity

- see Harrell ch. 4
- rules of thumb for inferential models with adequate power
  - e.g. $p < n/15$
  - effective n depends on data type (binary < small counts < continuous)
  - how does clustering/correlation in data change effective n?
- simplest if done **a priori**
  - data-driven choice of model complexity (e.g. by cross-validation), *while maintaining valid inference*, is **extremely** delicate ("post-selection inference")

Comparing models with reality

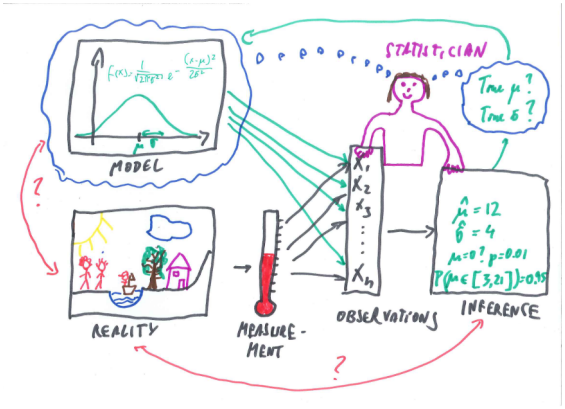

Figure 6: one view of models and reality, from Hennig (2022)

Model diagnostics

- all models make assumptions
- results *may* be sensitive to **misspecification**: bias, inefficiency, inflated/deflated type I error, poor coverage ...
- hypothesis tests (e.g. Shapiro-Wilk) are deprecated

Harvey Motulsky on CrossValidated:

> The question normality tests answer: Is there convincing evidence of any deviation from the Gaussian ideal? With moderately large real data sets, the answer is almost always yes.
>
> The question scientists often expect the normality test to answer: Do the data deviate enough from the Gaussian ideal to "forbid" use of a test that assumes a Gaussian distribution? Scientists often want the normality test to be the referee that decides when to abandon conventional (ANOVA, etc.) tests and instead analyze transformed data or use a rank-based nonparametric test or a resampling or bootstrap approach. For this purpose, normality tests are not very useful.

- "is there a statistically significant deviation from the model assumptions?" vs. "are the violations of the assumptions large enough to mess up my conclusions?" (**never** "are the data normally distributed?")
  - **data "too big"**: will reject assumptions even when it's OK
  - **data "too small"**: will fail to reject assumptions even when they're problematic (???)

- **two-stage testing** often has bad properties (H. Campbell and Dean 2014; Harlan Campbell 2021; Rochon, Gondan, and Kieser 2012; Zimmerman 2004; Shamsudheen and Hennig 2021)
- graphical diagnostics are often recommended
- but how do we judge whether deviations are too large? (Q-Q plot survey)
- **open question**

Untestable issues are often the worst problems

- biased/non-representative sampling (Meng 2018)
- lack of randomization
- not-at-random missingness
- problems with causal inference: unobserved confounders, etc.
- non-independence
  - pseudo-replication (Hurlbert 1984)
  - temporal/spatial structure

More on model diagnostics

- assumptions apply to **conditional** distribution of the response (not the marginal distribution, not the predictor variables)
- test assumptions in order of importance (George Box: "It is inappropriate to be concerned about mice when there are tigers abroad.")
  - bias/nonlinearity (residuals vs fitted plot)
  - heteroscedasticity (scale-location plot)
  - influential points/outliers (Cook's distance, leverage, etc.)
  - distributional assumptions (Q-Q plot)
  - ¿¿ correlated predictors ??
- `performance`, `DHARMa` packages: more later

References

Box, George E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* 71 (356): 791–99. https://doi.org/10.1080/01621459.1976.10480949.

Bryan, Jennifer. 2017. "Excuse Me, Do You Have a Moment to Talk about Version Control?" e3159v2. PeerJ Inc. https://doi.org/10.7287/peerj.preprints.3159v2.

Bryan, Jenny. 2017. "Project-Oriented Workflow." *Tidyverse*. https://www.tidyverse.org/blog/2017/12/workflow-vs-script/.

Bussola, Nicole, Alessia Marcolini, Valerio Maggio, Giuseppe Jurman, and Cesare Furlanello. 2020. "AI Slipping on Tiles: Data Leakage in Digital Pathology." arXiv. https://doi.org/10.48550/arXiv.1909.06539.

Campbell, Harlan. 2021. "The Consequences of Checking for Zero-Inflation and Overdispersion in the Analysis of Count Data." *Methods in Ecology and Evolution* 12 (4): 665–80. https://doi.org/10.1111/2041-210X.13559.

Campbell, H., and C. B. Dean. 2014. "The Consequences of Proportional Hazards Based Model Selection." *Statistics in Medicine* 33 (6): 1042–56. https://doi.org/10.1002/sim.6021.

Dushoff, Jonathan, Morgan P. Kain, and Benjamin M. Bolker. 2018. "I Can See Clearly Now: Reinterpreting Statistical Significance," October. https://arxiv.org/abs/1810.06387.

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.

Hastie, Trevor, Robert Tibshirani, and J. H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. https://hastie.su.domains/Papers/ESLII.pdf.

Heinze, Georg, Mark Baillie, Lara Lusa, Willi Sauerbrei, Carsten Oliver Schmidt, Frank E. Harrell, Marianne Huebner, and on behalf of TG2 and TG3 of the STRATOS initiative. 2024. "Regression Without Regrets –Initial Data Analysis Is a Prerequisite for Multivariable Regression." *BMC Medical Research Methodology* 24 (1): 178. https://doi.org/10.1186/s12874-024-02294-3.

Hennig, Christian. 2022. "Testing in Models That Are Not True." University of Bologna. https://www.wu.ac.at/fileadmin/wu/d/i/statmath/Research_Seminar/SS_2022/2022_04_Hennig.pdf.

Hurlbert, Stuart H. 1984. "Pseudoreplication and the Design of Ecological Field Experiments." *Ecological Monographs* 54 (2): 187–211. https://doi.org/10.2307/1942661.

Kapoor, Sayash, and Arvind Narayanan. 2023. "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science." *Patterns* 4 (9): 100804. https://doi.org/10.1016/j.patter.2023.100804.

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

Meng, Xiao-Li. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *Annals of Applied Statistics* 12 (2): 685–726. https://doi.org/10.1214/18-AOAS1161SF.

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

Quinn, Thomas P. 2021. "Stool Studies Don't Pass the Sniff Test: A Systematic Review of Human Gut Microbiome Research Suggests Widespread Misuse of Machine Learning." *arXiv:2107.03611* [*q-Bio*], July. https://arxiv.org/abs/2107.03611.

Rochon, Justine, Matthias Gondan, and Meinhard Kieser. 2012. "To Test or Not to Test: Preliminary Assessment of Normality When Comparing Two Independent Samples." *BMC*

Medical Research Methodology 12 (1): 81. https://doi.org/10.1186/1471-2288-12-81.

Schielzeth, Holger. 2010. "Simple Means to Improve the Interpretability of Regression Coefficients: Interpretation of Regression Coefficients." *Methods in Ecology and Evolution* 1 (2): 103–13. https://doi.org/10.1111/j.2041-210X.2010.00012.x.

Shamsudheen, M. Iqbal, and Christian Hennig. 2021. "Should We Test the Model Assumptions Before Running a Model-Based Test?" arXiv. https://arxiv.org/abs/1908.02218.

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.

Zimmerman, Donald W. 2004. "A Note on Preliminary Tests of Equality of Variances." *British Journal of Mathematical and Statistical Psychology* 57 (1): 173–81. https://doi.org/10.1348/000711004849222.