Review of linear models

2 Sep 2023

Basics

- assume $\mathbf{y} \sim \text{Normal}(\mathbf{X}\beta, \sigma)$[1]
- $\mathbf{X}$ is the *model matrix*, can be anything we want it to be
- the *Gauss-Markov theorem* (Wikipedia) makes weaker assumptions: $\mathbf{y} = \mathbf{X}\beta + \epsilon$; as long as ϵ is mean-zero, homoscedastic with finite variance, and uncorrelated ... then the OLS solution
$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$
is the BLUE (or MVUE).
- we'll embrace the assumptions (which are needed for inference!)

Inference

Model matrices

- everything has to be converted to $\mathbf{X}$ before we start
  - transformations
  - encoding of categorical variables: **contrasts**
  - interactions
  - basis expansions (e.g. polynomials)

Wilkinson-Rogers formulas

- Wilkinson and Rogers (1973), updated by Chambers and Hastie (1991, ch. 2)
- operators: `+`, `*`, `:`, `/`, `-`, `^`
- `I()`

[1] Notation-abuse warning ...

Contrasts

Marginality

- Venables (1998)
- 'type (X) sums of squares'
- scaling and centering (Schielzeth 2010)

Computation

- matrix decompositions (QR with pivoting)
- big problems: `biglm`, `speedglm`, `RcppEigen::fastLm`
  - optimized BLAS, kernel trick, etc.
  - memory vs speed vs robustness ...
  - p vs. n vs. many-small-regressions vs. ...

Diagnostics

- linearity,
- base R: `stats::plot.lm()`
- `performance::check_model()`
- `DHARMa` (`simulateResiduals(., plot = TRUE)`)

References

Chambers, J. M., and T. J. Hastie, eds. 1991. *Statistical Models in S*. 1st ed. Chapman; Hall/CRC.

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.

Venables, W. N. 1998. "Exegeses on Linear Models." In. 1998 International S-PLUS User Conference. Washington, DC. http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf.

Wilkinson, G. N., and C. E. Rogers. 1973. "Symbolic Description of Factorial Models for Analysis of Variance." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 22 (3): 392–99. https://doi.org/10.2307/2346786.