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
$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$
is the BLUE (or MVUE).
- we'll embrace the assumptions (which are needed for inference!)

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)

- operators: `+`, `*`, `:`, `/`, `-`, `^`
- `I()`

[1]Notation-abuse warning …

Contrasts

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

References

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.