Review of linear models

4 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!)

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. ...

Inference

- σ^2 (residual variance) is RSS$/(n-p)$
- The covariance matrix is $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$.
- Individual coefficients are t-distributed
- Linear combinations of coefficients are F-distributed
- Wald and likelihood ratio test comparisons are equivalent
  (but need to be careful about marginality)

[1] Notation-abuse warning ...

Model matrices

- model definition converted to **X** before we start
- **input variables** vs **predictor variables** (Schielzeth (2010), Gelman & Hill (2006), CV)
  - transformations
  - encoding of categorical variables: **contrasts**
  - interactions
  - basis expansions (e.g. polynomials)

Wilkinson-Rogers formulas

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

Contrasts

Marginality

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

Downstream methods

- prediction, effects plots
- uncertainty of predictions
- `emmeans`, `marginaleffects`, `effects`, `sjPlot` ...
- `tidy()`, `performance`, `insight`, etc. ...

Diagnostics

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

References

Chambers, J. M., & Hastie, T. J. (Eds.). (1991). *Statistical Models in S* (1st ed.). Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients: Interpretation of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103–113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

Venables, W. N. (1998). *Exegeses on Linear Models*. http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *22*(3), 392–399. https://doi.org/10.2307/2346786