Ridge regression and mixed models

16 Oct 2023

```
## it's nice to include packages at the top
## (and NOT automatically install them)
## try not to carry over packages you don't use
library(ggplot2); theme_set(theme_bw())
## diagnostics
library(performance)
library(DHARMa)
## downstream model evaluation
library(broom)
library(dotwhisker)
library(emmeans)
library(effects)
library(marginaleffects)
library(parameters)
## library(ggeffects)
```

Ridge in a nutshell

- **penalized** models: instead of minimizing $\text{SSQ} = \sum((\mathbf{y} - \mathbf{X}\beta)_i)^2$, minimize $\text{SSQ} + \lambda||\beta||_2$ (ridge)
- or + $||\beta||_1$ (lasso)
- optimize *bias-variance tradeoff*
- equivalent to imposing iid Gaussian priors on each element of β
- lasso (and elastic net, which is a convex combination of L2 and L1 penalties) are popular because they **induce sparsity**
  - *likelihood surfaces* are non-convex with cusps at zero

    - optimization with non-convex surfaces is a nuisance because it makes the basic optimization problem nonlinear; we need to use a different algorithm (coordinate descent/soft thresholding); can't use *only* linear algebra

- can generalize from penalized LM to penalized GLM

Andrew Gelman on variable selection

> Variable selection (that is, setting some coefficients to be exactly zero) can be useful for various reasons, including: * It's a simple form of regularization. * It can reduce costs in future data collection. Variable selection can be fine as a means to an end. Problems can arise if it's taken too seriously, for example as an attempt to discover a purported parsimonious true model.

Choosing penalty strength

- typically by *cross-validation*
- leave-one-out (LOOCV) vs k-fold

Practical points

- Predictors **must** be standardized
- Intercept should usually be unpenalized

Ridge vs lasso

- In practice people just try both (or elastic net)
- Conjecture: whether ridge or lasso is a better *predictive* model in a particular case depends on the *effect size spectrum*

Ridge by data augmentation

- set

$$\mathbf{B} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{pmatrix}$$

- and $\mathbf{y}^* = (\mathbf{y}\ \mathbf{0})$
- so that $\mathbf{B}^\top \mathbf{B} = \mathbf{X}^\top \mathbf{X} + \lambda I$ and the residual sum of squares is unchanged

From ridge to mixed models

- what if we say

$$
\begin{aligned}
\mathbf{y} &\sim \text{Normal}(\mathbf{X}\beta, \sigma^2) \\
\beta &\sim \text{MVN}(\mathbf{0}, \sigma_g^2 \mathbf{I})
\end{aligned}
$$

?

i.e. treat this as an *empirical Bayesian* problem (we estimate the β values, but do not put a prior on σ^2 or a hyperprior on σ_g^2 $(= 1/\lambda)$

References