There are many alternatives to using least squares to fit. Three of the most important classes are:

- **subset selection**: we first identify a subset of the predictors which are related to the response, then use least squares on the reduced set of variables
- **shrinkage**: fit a model involving *all* p predictors, but shrink the estimated coefficients towards zero as a regularisation, with the effect of reducing variance. Note that some methods could also allow some of the parameters to be zero, thus performing variable selection too
- **dimensionality reduction**: *projecting* the p predictors into an m-dimensional space, where m << p

The two best-known techniques for shrinking the regression coefficients towards zero are **ridge regression** and the **lasso**.

Ridge regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

The lambda term is called *shrinkage penalty*.

- for lambda=0, we have plain least squares
- for lambda->inf, we have smaller and smaller betas
- we do have a different set of betas for every lambda. Choosing the right lambda is critical

NOTE: the index j starts from 1 => we do not shrink the intercept beta_0.

(see R code)

The standard least squares coefficient estimates are *scale equivariant*: regardless of how the jth predictor is scaled, Xj*betaHat_j will remain the same.

In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant. Therefore, it is best to apply ridge regression *after standardizing* the predictors, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

As λ increases, the flexibility of the ridge regression fit decreases,leading to decreased variance but increased bias.

LIMIT: ridge regression will always include all predictors into the model

Lasso

The lasso coefficients minimise the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \ = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

(note the L1 norm instead of L2)
Lasso yields *sparse* models

We can interpret ridge regression and the lasso as computationally feasible alternatives to best subset selection that replace the intractable form of the budget with forms that are much easier to solve.

Additional links:

- https://gerardnico.com/wiki/lang/r/ridge_lasso