Linear Regression with Regularization

- Ridge regression

$$\hat{\boldsymbol{\beta}} = \mathsf{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2.$$

- Lasso

$$\hat{\boldsymbol{\beta}} = \mathsf{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda |\boldsymbol{\beta}|, \tag{1}$$

- Subset selection

$$\hat{\boldsymbol{\beta}} = \mathsf{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_0,$$

  which with a proper choice of λ gives rise to AIC, BIC, or Mallow's C_p when σ^2 is known or estimated by a plug-in.

- $\|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{p} \beta_j^2$, $|\boldsymbol{\beta}| = \sum_{j=1}^{p} |\beta_j|$, $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathbf{1}_{\{\beta_j \neq 0\}}$.

Note that the penalty or regularization terms are not invariant with respect to any location/scale change of the predictors, so we usually

- center and scale the columns of the design matrix $\mathbf{X}$ such that they have mean zero and unit sample variance, and
- center $\mathbf{y}$, so the intercept is suppressed (why).

Some packages in R (e.g., `glmnet`) handles the centering and scaling automatically: they apply the transformation before running the algorithm, and then transform the obtained coefficients back to the original scale and add back the intercept.

How to compute the intercept?

$$Y - \bar{y} = \hat{\beta}_1(X_1 - \bar{\mathbf{x}}_1) + \hat{\beta}_2(X_2 - \bar{\mathbf{x}}_2) + \cdots + \hat{\beta}_p(X_p - \bar{\mathbf{x}}_p).$$

$$\Longrightarrow \hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{\mathbf{x}}_j.$$