Project: Cross-validation for model selection

Diego Urdapilleta de la Parra

March 3, 2025

Background

Setup and preliminary definitions

Let $\mathcal{D}_n := \{(y_i, \boldsymbol{x}_i) : i \in [n]\}$ be a set of independent data points drawn from a distribution $\mathbb{P}_{y,\boldsymbol{x}}$ for $(y, \boldsymbol{x}) \in \mathbb{R}^{1+p}$. We treat the $\boldsymbol{x}_i$ as predictors of the outcome y_i, and we assume a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_n]^\top$, and $\boldsymbol{e}$ is a mean-zero random vector with $\mathrm{Cov}\left(\boldsymbol{e}\right) = \sigma_2 \boldsymbol{I}_n$.

In the context of competing models

Define

$$\mathcal{L}_n\left(\alpha, \mathcal{D}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^\top\boldsymbol{\beta} + e_i - \boldsymbol{x}_{i\alpha}^\top\hat{\boldsymbol{\beta}}_\alpha\right)^2$$

Lemma 1

$$\mathbb{E}\left[\mathcal{L}_n\left(\alpha, \mathcal{D}_n\right) \,\middle|\, \boldsymbol{X}\right] = \sigma^2 + \frac{1}{n}d_\alpha\sigma^2 + \frac{1}{n}\left|\left|M_\alpha\boldsymbol{X}\boldsymbol{\beta}\right|\right|^2$$

Proof.

$$\mathbb{E}\left[\mathcal{L}_n\left(\left(\right)\alpha, \mathcal{D}_n\right), \boldsymbol{y}, \boldsymbol{X}\right] =$$

□