Project: Cross-validation for model selection (rough draft)

Diego Urdapilleta de la Parra

March 4, 2025

1 Background

1.1 Setup and preliminary definitions

Assumptions

$$\begin{aligned}
&\mathbf{H1}: \quad && \liminf_{n\to\infty} \frac{1}{n} \left\| M_\alpha \boldsymbol{X}\boldsymbol{\beta} \right\|^2 > 0 \text{ for all } \alpha \in \mathcal{A} \\
&\mathbf{H2}: \quad && \boldsymbol{X}^\top \boldsymbol{X} = O(n) \quad \text{and} \quad \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = O\left(n^{-1}\right) \\
&\mathbf{H3}: \quad && \lim_{n\to\infty} \max_{i\le n} h_{ii,\alpha} = 0
\end{aligned}$$

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 $\operatorname{Cov}(\boldsymbol{e}) = \sigma_2 \boldsymbol{I}_n$.

Definition 1.1

We denote the average squared error of $\hat{m}_\alpha$ by

$$\mathcal{L}_n(\alpha) := \frac{1}{n} \left\| \mathbb{E}[y \mid \boldsymbol{X}] - \hat{m}_\alpha(\boldsymbol{X}) \right\|^2.$$

Additionally, we write $R_n(\alpha) := \mathbb{E}[\mathcal{L}_n(\alpha) \mid \boldsymbol{X}]$.

Proposition 1.1

If we assume a linear model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}$, then

$$\mathcal{L}_n(\alpha) = \frac{1}{n}\|H_\alpha \boldsymbol{e}\|^2 + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \text{and} \quad R_n(\alpha) = \frac{1}{n}\sigma^2 p(\alpha) + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

where $H_\alpha = \boldsymbol{X}_\alpha\left(\boldsymbol{X}_\alpha^\top \boldsymbol{X}_\alpha\right)^{-1}\boldsymbol{X}_\alpha^\top$ and $M_\alpha = I_n - H_\alpha$.

Proof. First, we have that

$$\begin{aligned}
\|\mathbb{E}[\boldsymbol{y} \mid \boldsymbol{X}] - \hat{m}_\alpha(\boldsymbol{X})\|^2 &= \left\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\right\|^2 \\
&= \|\boldsymbol{X}\boldsymbol{\beta} - H_\alpha(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e})\|^2 \\
&= \|M_\alpha \boldsymbol{X}\boldsymbol{\beta} - H_\alpha \boldsymbol{e}\|^2.
\end{aligned}$$

Notice that $M_\alpha \boldsymbol{X}\boldsymbol{\beta}$ and $H_\alpha \boldsymbol{e}$ are orthogonal:

$$\boldsymbol{e}^\top H_\alpha M_\alpha \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{e}^\top H_\alpha(I_n - H_\alpha)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{e}^\top H_\alpha \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{e}^\top H_\alpha \boldsymbol{X}\boldsymbol{\beta} = 0.$$

Hence, by the Pythagorean theorem, the first part is satisfied.

For the second part, we note that $\mathbb{E}\left[\|H_\alpha \boldsymbol{e}\|^2 \mid \boldsymbol{X}\right] = \sigma^2 p(\alpha)$ by the "trace trick," where $p(\alpha)$ denotes the size of model α. ☐

Proposition 1.2

Suppose that the set of correct candidate models $\mathcal{A}_c \subset \mathcal{A}$ is non-empty, and let α^* be the smallest correct model in $\mathcal{A}_c$. Then, α^* minimizes $R_n(\alpha)$ over $\alpha \in \mathcal{A}$.

Proof. Let $\alpha \in \mathcal{A}$ be arbitrary and suppose that $\alpha \in \mathcal{A}_c$. Then, $\boldsymbol{X}_\alpha\boldsymbol{\beta}_\alpha = \boldsymbol{X}\boldsymbol{\beta}$ and $p(\alpha^*) \leq p(\alpha)$. Thus,

$$\begin{aligned}
R_n(\alpha) &= \frac{1}{n}\sigma^2 p(\alpha) + \frac{1}{n}\|M_\alpha \boldsymbol{X}\boldsymbol{\beta}\|^2 \\
&= \frac{1}{n}\sigma^2 p(\alpha) + \frac{1}{n}\underbrace{\|M_\alpha \boldsymbol{X}_\alpha\boldsymbol{\beta}_\alpha\|^2}_{0} \\
&= \frac{1}{n}\sigma^2 p(\alpha) \;\geq\; \frac{1}{n}\sigma^2 p(\alpha^*) = R_n(\alpha^*).
\end{aligned}$$

Now suppose that $\alpha \in \mathcal{A}_w$. If $p(\alpha) \geq p(\alpha^*)$, the result follows by assumption H1. If $p(\alpha) \geq p(\alpha^*)$, then ... MISSING. ☐

2 Leave-One-Out CV

Definition 2.1

The LOOCV estimator of $R_n(\alpha)$ is

$$\hat{R}_n^{(1)}(\alpha) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{x}_{i\alpha}^\top \hat{\boldsymbol{\beta}}_\alpha}{1-h_{ii,\alpha}}\right)^2$$

Lemma 2.1 (Shao, 1993)

$$\hat{R}_n^{(1)}(\alpha) = \begin{cases} \frac{1}{n}\|\boldsymbol{e}\|^2 + \frac{2}{n}\sigma^2 p(\alpha) - \|H_\alpha \boldsymbol{e}\|^2 + o_{\mathbb{P}}(n^{-1}) & \text{if } \alpha \in \mathcal{A}_c \\ R_n(\alpha) + o_{\mathbb{P}}(1) & \text{otherwise} \end{cases}$$

A corollary of this Lemma is that $\hat{R}_n^{(1)}(\alpha) \to_{\mathbb{P}} \sigma^2 \leftarrow_{\mathbb{P}} R_n(\alpha)$.

Proposition 2.2 (Shao, 1993)

Let $\hat{\alpha}^{(1)}$ be the model minimizing $\hat{R}_n^{(1)}(\alpha)$.

1. Under H1, H2, and H3,
$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\alpha}^{(1)} \in \mathcal{A}_w\right) = 0.$$
2. If $p(\alpha) \neq p$,
$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\alpha}^{(1)} = \alpha^*\right) \neq 1$$
3. For $\alpha \in \mathcal{A}_c$ with $\alpha \neq \alpha^*$,
$$\mathbb{P}\left(\hat{R}_n^{(1)}(\alpha) \leq \hat{R}_n^{(1)}(\alpha^*)\right) = \mathbb{P}\left(2\left(p(\alpha) - p(\alpha^*)\right)\sigma^2 < \boldsymbol{e}^\top (H_\alpha - H_{\alpha^*})\boldsymbol{e}\right) + o_{\mathbb{P}}(1).$$
If $\boldsymbol{e} \sim \mathbb{N}(0_n, \sigma^2 I_n)$,
$$\mathbb{P}\left(\hat{R}_n^{(1)}(\alpha) \leq \hat{R}_n^{(1)}(\alpha^*)\right) = \mathbb{P}\left(2k < \chi^2(k)\right) + o_{\mathbb{P}}(1)$$
for $k = p(\alpha) - p(\alpha^*)$.

Corollary 2.3

LOOCV overfits with non-zero probability asymptotically.