In-sample prediction and Mallow's C_p

Consider a linear regression model with p predictors (let's ignore the intercept in this note).

- Index all possible variable subsets by a p-dimensional binary vector (totally 2^p subsets or models):
$$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^t \in \{0,1\}^p.$$
Especially, $\boldsymbol{\gamma} = (1, 1, \ldots, 1)$ denotes the biggest (full) model that includes all the predictors, and $\boldsymbol{\gamma} = (0, 0, \cdots, 0)$ denotes the smallest (null) model that does not include any predictors.

- For a variable set $\boldsymbol{\gamma}$, define $p_{\boldsymbol{\gamma}} = \sum_j \gamma_j$ to denote the number of variables included in this set, and use $\mathbf{X}_{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ to denote the corresponding $n \times p_{\boldsymbol{\gamma}}$ design matrix and $p_{\boldsymbol{\gamma}}$-dim LS regression parameter, respectively.

In-sample prediction

The so-called *in-sample prediction error* measures prediction errors at the n sample points $\mathbf{x}_i$'s. For a model $\boldsymbol{\gamma}$, the error is defined to be

$$R(\boldsymbol{\gamma}) = \mathbb{E}\|\mathbf{y}^* - \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2, \tag{1}$$

where

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}, \quad \mathbf{X}_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y},$$

and $\mathbf{y}^*$ is a set of imaginary, new data points observed at $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ which are independent of the training data $\mathbf{y}$.

The n-by-n matrix $\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T$ is known as the projection matrix or hat matrix. It is symmetric and idempotent with $\text{tr}(\mathbf{H}_{\boldsymbol{\gamma}}) = p_{\boldsymbol{\gamma}}$.

The expectation in (1) is taken with respect to the true distribution over $\mathbf{y}$ and $\mathbf{y}^*$. Here is our assumption on the true data generating process:

$$\mathbf{y}_{n\times 1},\ \mathbf{y^*}_{n\times 1} \text{ i.i.d.} \sim \mathcal{N}_n\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n\right). \tag{2}$$

Or equivalently, assume

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\mu} + \mathbf{e}, \\ \mathbf{y}^* &= \boldsymbol{\mu} + \mathbf{e}^* \\ \mathbf{e}_{n\times 1},\ \mathbf{e^*}_{n\times 1} \text{ i.i.d.} &\sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right). \end{aligned}$$

Note that *1)* we do not model the randomness of the X features and the design matrix $\mathbf{X}$ is assumed to be given (the usual setup in statistical analysis for linear models); *2)* we do not need to assume the mean vector $\boldsymbol{\mu}$ can be expressed as a linear combination of the design matrix $\mathbf{X}$, in other words, whether the true model is a linear model or not does not affect our analysis.

Next we decompose the prediction error into three components.

$$
\begin{aligned}
R(\boldsymbol{\gamma}) &= \mathbb{E}\|\mathbf{y}^* - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\
&= \mathbb{E}\|(\mathbf{y}^* - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\
&= \mathbb{E}\|\mathbf{y}^* - \boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \mathbb{E}\|\mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\
&= I + II + III.
\end{aligned} \tag{3}
$$

The symbol $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\mathbf{X}_\gamma^t \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^t \boldsymbol{\mu}$ is defined to be the best choice of LS coefficients for model $\boldsymbol{\gamma}$ if we knew the true mean vector $\boldsymbol{\mu}$. It is easy to show that $\mathbb{E}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

Note that all the cross-product terms are equal to zero:

$$
\begin{aligned}
\mathbb{E}(\mathbf{y}^* - \boldsymbol{\mu})^t(\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}}) &= (\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}})^t \; \mathbb{E}(\mathbf{y}^* - \boldsymbol{\mu}) = (\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}})^t \mathbf{0} = 0. \\
\mathbb{E}(\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}})^t(\mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) &= (\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}})^t \; \mathbb{E}(\mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = 0 \\
\mathbb{E}\Big[(\mathbf{y}^* - \boldsymbol{\mu})^t(\mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\Big] &= \Big[\mathbb{E}(\mathbf{y}^* - \boldsymbol{\mu})\Big]^t \Big[\mathbb{E}(\mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\Big] = 0
\end{aligned}
$$

where at the last equality we use the fact that $\mathbf{y}^*$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ (depending on $\mathbf{y}$) are uncorrelated.

Let us examine each term in (3)

- The 1st term: the unavoidable error that you would encounter even if you know the true parameter $\boldsymbol{\beta}$:

$$I = \|\mathbf{y}^* - \boldsymbol{\mu}\|^2 = \mathbb{E}\|\mathbf{e}^*\|^2 = n\sigma^2.$$

- The 2nd term: As we explained before, $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = (\mathbf{X}_\gamma^t \mathbf{X}_\gamma)^{-1} \mathbf{X}_\gamma^t \boldsymbol{\mu}$ is the solution of the following LS problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{p_\gamma}} \|\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\alpha}\|^2 = \|\boldsymbol{\mu} - \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 = II.$$

  The bias will be zero, if $\mu = \mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}}$ (this would happen if the true model is a linear model and $\boldsymbol{\gamma}$ contains all the true predictors). The bias will not be zero, e.g., if the model $\boldsymbol{\gamma}$ misses any true predictors.

- The 3rd term: the variance of model $\boldsymbol{\gamma}$ (due to estimating $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$). Note that $\mathbf{X}_\gamma \boldsymbol{\beta}_{\boldsymbol{\gamma}} = \mathbf{H}_\gamma \boldsymbol{\mu}$ and $\mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \mathbf{H}_\gamma \mathbf{y}$. Then

$$
\begin{aligned}
III &= \mathbb{E}\|\mathbf{H}_\gamma \boldsymbol{\mu} - \mathbf{H}_\gamma \mathbf{y}\|^2 \\
&= \mathbb{E}\|\mathbf{H}_\gamma (\mathbf{y} - \boldsymbol{\mu})\|^2 = \sigma^2 \mathrm{tr}(\mathbf{H}_\gamma) = \sigma^2 p_\gamma.
\end{aligned}
$$

Mallow's C_p

In practice, we do not know the true model, i.e., we cannot calculate the 2nd term (the bias). We try to get that information from the RSS from model $\boldsymbol{\gamma}$. Next let's look at the expected $\text{RSS}_{\boldsymbol{\gamma}}$, and do a similar decomposition.

$$\begin{aligned}
\mathbb{E}[\text{RSS}_{\boldsymbol{\gamma}}] &= \mathbb{E}_{\mathbf{y}}\|\mathbf{y} - \mathbf{X}_{\gamma}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\
&= \mathbb{E}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\beta} - \mathbf{X}_{\gamma}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{X}_{\gamma}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\gamma}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\
&= \mathbb{E}\|\mathbf{y} - \boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu} - \mathbf{X}_{\gamma}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2 + \mathbb{E}\|\mathbf{X}_{\gamma}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbf{X}_{\gamma}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\|^2 \\
&\quad -2\mathbb{E}(\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{X}_{\gamma}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \mathbf{X}_{\gamma}\boldsymbol{\beta}_{\boldsymbol{\gamma}}). \qquad (4)
\end{aligned}$$

The first 3 terms are the same as the ones in (3). But now we have a cross-product term which does not appear in our derivation for prediction. If we replace $\mathbf{y}$ by $\mathbf{y}^*$ in (4), since the new test data $\mathbf{y}^*$ is independent of the training data $\mathbf{y}$, the cross-product term is zero (they are uncorrelated). But for RSS, the data $\mathbf{y}$ is used both for evaluation and for learning (it's used twice), so the cross-product term will not disappear.

The cross-product term (the last term) in (4) is equal to

$$\mathbb{E}(\mathbf{y} - \boldsymbol{\mu})^T(\mathbf{H}_{\gamma}\mathbf{y} - \mathbf{H}_{\gamma}\boldsymbol{\mu}) = \sigma^2 \text{tr}(\mathbf{H}_{\gamma}) = \sigma^2 p_{\boldsymbol{\gamma}}$$

So

$$R(\boldsymbol{\gamma}) \approx \text{RSS} + 2p_{\boldsymbol{\gamma}}\sigma^2,$$

which gives rise to Mallow's C_p: $\text{RSS}_{\boldsymbol{\gamma}} + 2p_{\boldsymbol{\gamma}}\hat{\sigma}^2$, where we replace σ^2 by an estimate from the full model.