

We are given a data set  $\{y_i, \mathbf{x}_i\}_{i=1}^n$  where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^\top$ 's are  $k \times 1$  vectors and  $y_i$ 's are scalars. It is known that the **Data Generating Process (DGP)** is

$$y_i = \mu(\mathbf{x}_i) + \epsilon_i$$

where  $\epsilon_i$  is a random variable with mean zero and variance  $\sigma^2$ . That is, for each input  $\mathbf{x}_i$ , an output  $y_i$  is produced with an **independent and identically distributed (iid)** error  $\epsilon_i$ . However,  $\mu$  is an unknown function and it might only utilize a subset of the  $k$  inputs. We represent the process compactly as  $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$  where  $\mathbf{y} = (y_1, \dots, y_n)^\top$ ,  $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n))^\top$ , and  $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$ .

Our job is simple: predict  $y_i$  with  $\mathbf{x}_i$  using a linear model. However, how do we know which of  $k$  the inputs of  $\mathbf{x}_i$  should we put in our model? That is, we want to find a way to measure how the good the prediction of a specific model would be.

Let  $\mathcal{A} \subseteq \{1, \dots, k\}$  with  $|\mathcal{A}| = p$  denote a subset of the indices of size  $p$  and let  $\mathbf{X}_{\mathcal{A}}$  denote the corresponding data matrix  $(\mathbf{x}_{1\mathcal{A}}, \dots, \mathbf{x}_{n\mathcal{A}})^\top$ . That is,  $\mathcal{A}$  denotes the subset of the  $k$  independent variables we choose to put in our model. The standard **Ordinary Least Square (OLS)** estimator yields the estimator  $\hat{\boldsymbol{\beta}}_{\mathcal{A}} = (\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{X}_{\mathcal{A}}^\top \mathbf{y}$ . An intuitive way of measuring prediction quality is to consider the expected sum of square errors:

$$\begin{aligned} \mathbf{E} \left[ \sum_{i=1}^n (y_i - \mathbf{x}_{i\mathcal{A}}^\top \hat{\boldsymbol{\beta}}_{\mathcal{A}})^2 \right] &= \mathbf{E}(\mathbf{y} - \mathbf{X}_{\mathcal{A}} \hat{\boldsymbol{\beta}}_{\mathcal{A}})^\top (\mathbf{y} - \mathbf{X}_{\mathcal{A}} \hat{\boldsymbol{\beta}}_{\mathcal{A}}) \\ &= n\sigma^2 - p\sigma^2 + \boldsymbol{\mu}^\top (\mathbf{I}_n - \mathbf{P}_{\mathcal{A}}) \boldsymbol{\mu} \end{aligned} \quad (\text{in})$$

where  $\mathbf{P}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} (\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{X}_{\mathcal{A}}^\top$  and  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Notice the term  $-p\sigma^2$ . This term suggests that the prediction error decreases as  $p$ , the size of  $\mathcal{A}$ , increases. That is, we can keep adding inputs from the original  $k$  independent variables to the linear model and the square error will decrease! Therefore, this expected sum of squares error is not a good measure for how good the model will perform.

However, notice this this is only the case when we are doing “in-sample” prediction, i.e., evaluating sum of squares error with the data set that is used to produce the estimate  $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ . We can consider calculating the prediction error with a *hypothetical out-sample data set*, that is, a data set  $\{y_i^{\text{out}}, \mathbf{x}_i\}_{i=1}^n$  where

$$y_i^{\text{out}} = \mu(\mathbf{x}_i) + \epsilon_i^{\text{out}}.$$

This hypothetical data set is essentially “a set of regenerated  $y_i$ ’s with the same  $\mathbf{x}_i$ ’s.” Using the new data set, we can compute the “out-sample” prediction error associated with the “in-sample” estimate  $\hat{\beta}_{\mathcal{A}}$ :

$$\begin{aligned} \mathbf{E} \left[ \sum_{i=1}^n (y_i^{\text{out}} - \mathbf{x}_{i\mathcal{A}}^\top \hat{\beta}_{\mathcal{A}})^2 \right] &= \mathbf{E} (\mathbf{y}^{\text{out}} - \mathbf{X}_{\mathcal{A}} \hat{\beta})^\top (\mathbf{y}^{\text{out}} - \mathbf{X}_{\mathcal{A}} \hat{\beta}_{\mathcal{A}}) \\ &= n\sigma^2 + p\sigma^2 + \boldsymbol{\mu}^\top (\mathbf{I}_n - \mathbf{P}_{\mathcal{A}}) \boldsymbol{\mu}. \end{aligned} \quad (\text{out})$$

Notice how the out-sample prediction error increases as  $p$ , number of independent variables in our model, increases. Hence, out-sample prediction error is a much better criterion for evaluating the fitness of a model.

Now the practical question: How can we calculate the “out-sample prediction error” when we only observe one data set? The trick is to approximate the out-sample prediction error with the in-sample prediction error. In fact, **(in)** and **(out)** are related by the simple equation

$$\text{out} = \text{in} + 2p\sigma^2. \quad (1)$$

The term  $2p\sigma^2$  can be viewed as an error correction term to **(in)**. We can replace  $\sigma^2$  by some estimator  $\hat{\sigma}^2$  to obtain an estimate of **(out)**. And that’s basically it!

\* \* \*

Formally,  $C_p$  is defined as follows: Suppose we have data  $\{y_i, \mathbf{x}_i\}_{i=1}^n$  as before, and we pick  $p$  of the  $k$  exogenous variables from  $\mathbf{x}_i$  to calculate the linear model coefficients  $\beta$ , denoted by  $\hat{\beta}_{\mathcal{A}}$ . The Mallows’  $C_p$  for that choice of  $p$  variables is defined by

$$C_p := \frac{1}{n} \left( \sum_{i=1}^n (y_i - \mathbf{x}_{i\mathcal{A}}^\top \hat{\beta}_{\mathcal{A}})^2 + 2p\hat{\sigma}^2 \right) \quad (2)$$

It is clear that **(2)** is simply **(1)** divided by  $n$ . The  $C_p$  values for different choices of  $\mathcal{A}$  tell us how the fitness of these models differ. The choice of  $\mathcal{A}$  with the smallest  $C_p$  is the most preferable. ■

## Acronyms

<b>DGP</b>	Data Generating Process. <b>1</b>
<b>iid</b>	independent and identically distributed. <b>1</b>
<b>OLS</b>	Ordinary Least Square. <b>1</b>