Mallows's C_p 2023-12-12

jessekelighine.com

We are given a data set $\{x_i, y_i\}_{i=1}^n$ where x_i's are $k \times 1$ vectors and y_i's are scalars. It is known that the data generating process (DGP) is

$$y_i = \mu(x_i) + e_i$$

where

$$\mu(x_i) = x_i^\mathsf{T} \beta$$

with $\mathbf{E}_{x_i}(e_i) = 0$ and $\text{Var}(e_i \mid x_i) = \sigma^2$. [1] Compactly, we can write the process as $Y = \mu(X) + E$ where $Y = [y_1, ..., y_n]^\mathsf{T}$, $X = [x_1, ..., x_n]^\mathsf{T}$, and $E = [e_1, ..., e_n]^\mathsf{T}$.

Our job is simple: to predict y_i given x_i using a linear model, i.e., to assess the "fitness" of the model. However, how do we know which of k the exogenous variables in x_i should we choose to put in our model? We want to find a way to measure how the good the prediction of a specific model would be.

The standard OLS estimator yields the estimator $\hat{\beta} = (X'X)^{-1}X'Y$. An intuitive way of measuring prediction quality is to consider the expected sum of square errors:

$$\begin{aligned} \mathbf{E}_X \left[\sum_{i=1}^n (y_i - x_i^\mathsf{T} \hat{\beta})^2 \right] &= \mathbf{E}_X (Y - X\hat{\beta})^\mathsf{T} (Y - X\hat{\beta}) \\ &= n\sigma^2 - k\sigma^2 + \mu^\mathsf{T}(I_n - P)\mu \qquad \text{(in)} \end{aligned}$$

where $P = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$. Notice the term $-k\sigma^2$. This term suggests that the prediction error decreases as k, number of exogenous variables, increases. That is, we can keep adding unrelated exogenous variables to the linear model and the prediction error will decrease! Thus, this prediction error is not a good measure for how good the model is.

However, notice this this is only the case when we are doing "in-sample" prediction, i.e., evaluating prediction error with the data set that is used to produce $\hat{\beta}$. We can consider calculating the prediction error using a hypothetical new data set with the same x_i's but with different y_i's, denoted by y_i^{out}, generated according to the data generating process. Using the new data set $\{x_i, y_i^{\text{out}}\}$, we can compute the "out-sample" prediction error:

$$\begin{aligned} \mathbf{E}_X \left[\sum_{i=1}^n (y_i^{\text{out}} - x_i^\mathsf{T} \hat{\beta})^2 \right] &= \mathbf{E}_X (Y^{\text{out}} - X\hat{\beta})^\mathsf{T} (Y^{\text{out}} - X\hat{\beta}) \\ &= n\sigma^2 + k\sigma^2 + \mu^\mathsf{T}(I_n - P)\mu. \qquad \text{(out)} \end{aligned}$$

[1] $\mathbf{E}_X(\cdot)$ denotes $\mathbf{E}(\cdot \mid X)$.

It is clear that the out-sample prediction error increases as k increases. Hence, out-sample prediction error is a much better criterion for evaluating the fitness of a model.

Now the practical question is: 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

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

The term $2k\sigma^2$ can be viewed as an error correction term to (in). We can replace σ^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 $\{y_i, x_i\}$ as before, and we pick p of the k exogenous variables from x_i to calculate the linear model coefficients $\hat{\beta}$, denoted by $\hat{\beta}_p$. The C_p for that choice of p variables is defined by

$$C_p \coloneqq \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - x_i^{\mathsf{T}}\hat{\beta}_p)^2 + 2p\hat{\sigma}^2\right) \tag{2}$$

It is clear that (2) is simply (1) divided by n. The C_p values for different choices of p tell us how the fitness of these models differ. The choice of p with the smallest C_p is the most preferable.