

Cross-Validation methods in regression problems

Il metodo di cross-validation nei problemi di regressione¹

Daniela Cocchi, Meri Raggi
Dipartimento di Scienze Statistiche
Università di Bologna
cocchi@stat.unibo.it, raggi@stat.unibo.it

Riassunto: L'approccio unitario qui proposto permette di estendere direttamente gli sviluppi teorici della cross-validation noti dalla letteratura per il modello lineare classico ad una classe di modelli molto più ampia. In particolare sono derivati i risultati riguardanti il modello di Fay and Herriot, che appartiene alla classe dei modelli lineari misti.

Keywords: Balanced Incomplete Block design, cross-validation, data splitting, design matrix.

1. Introduction

In this work we discuss the cross-validation method, addressing our attention to asymptotic properties in linear regression problems. The main contributions of the literature, like Stone (1978), Picard and Cook (1984) and Shao (1993) are developed without using matrix algebra. We introduce a compact matrix notation allowing us to derive the main results in a way which provides straightforward calculations together with the continuous monitoring of the contribution of the different elements to the solution.

2. Variable selection in the linear model

Under the linear framework the model selection problem coincides with predictor selection. Usually this problem is faced within a “perfect” setting, where it is assumed that an adequate data set is available, which includes all relevant covariates and, perhaps, some others extraneous to the phenomenon investigated. Consider the simplest linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y}$ denotes the n -dimension vector of observations of the dependent variable, $\mathbf{X}$ is the matrix ($n \times p$) containing the values of independent variables, $\boldsymbol{\beta}$ is the p -dimension vector of unknown parameters and $\boldsymbol{\varepsilon}$ is the vector of random errors, where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$. Model (1) has potentially p independent variables from which a subset of variables is to be selected into a reduced model. We manage the variable selection by introducing a design matrix $\mathbf{M}$:

$$\mathbf{M} = (\mathbf{e}_{M_1}, \dots, \mathbf{e}_{M_j}, \dots, \mathbf{e}_{M_\alpha})$$

where $\mathbf{e}_{M_j}$ is the M_j -th column of the identity matrix $\mathbf{I}_p$ and $\alpha \leq p$ is the number of variables selected in the model. Then the reduced model with α covariates, $\mathcal{M}_\alpha$, can be

¹The research leading to this paper has been partially funded by a 2000 grant (project No.26; sector: Economics and Statistics) for research projects of national interest by MIUR.

written as:

$$\mathbf{y} = \mathbf{XMM}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\alpha$$

where $\mathbf{M}'\boldsymbol{\beta}$ can be estimated, by OLS, as: $\mathbf{b}_\alpha = (\mathbf{M}'\mathbf{X}'\mathbf{XM})^{-1}\mathbf{M}'\mathbf{X}'\mathbf{y}$. The dimension α of the reduced model is equal to the number of units in the matrix $\mathbf{M}$, *i.e.* $tr(\mathbf{M}'\mathbf{M})$. Let $\mathbf{i}_\alpha$ and $\mathbf{i}_{\alpha^*}$ be the vectors composed by α and α^* units respectively. Two p -dimension vectors can be constructed as:

$$\begin{aligned}\mathbf{c}_\alpha &= \mathbf{M}_\alpha \cdot \mathbf{i}_\alpha \\ \mathbf{c}_{\alpha^*} &= \mathbf{M}_* \cdot \mathbf{i}_{\alpha^*}\end{aligned}$$

and a summary vector $\mathbf{c}$ as: $\mathbf{c} = \mathbf{c}_{\alpha^*} - \mathbf{c}_\alpha = (c_1, \dots, c_i, \dots, c_p)'$. If $\exists c_i > 0$ for $i = 1, \dots, p$ then $\mathcal{M}_\alpha$ is incorrect; if $c_i \leq 0$ for all $i = 1, \dots, p$ then $\mathcal{M}_\alpha$ is correct, if $c_i = 0$ for all $i = 1, \dots, p$ then $\mathcal{M}_\alpha$ is the optimal model, $\mathcal{M}_\alpha = \mathcal{M}_*$. Shao (1993) defines two categories of models: category I which contains incorrect models; on the contrary, category II contains models which include all relevant variables, these models are called “correct”, but may be inefficient because they may contain an unnecessary large number of variables. The smallest correct model, that arises from a particular $\mathbf{M}_*$ matrix, is $\mathcal{M}_*$, the optimal one possessing the smallest expected prediction error with respect to any other model.

3. The cross validation method

By means of an initial data set, a model $\mathcal{M}_\alpha$ which is able to predict the values y_i^* is estimated. If a response vector $\mathbf{y}^*$, different from the one used for model estimation, and the corresponding vector of covariates $\mathbf{x}_i$ are available, then the average squared prediction error is:

$$n^{-1} \sum_{i=1}^n (y_i^* - \hat{y}_i^*)^2 = n^{-1} \sum_{i=1}^n \left(\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i - \mathbf{x}_i' \mathbf{M} (\mathbf{M}' \mathbf{X}' \mathbf{X} \mathbf{M})^{-1} \mathbf{M}' \mathbf{X}' \mathbf{y} \right)^2. \quad (2)$$

Given $\mathbf{y}$, the conditional expected squared prediction error is:

$$E \left(n^{-1} \sum_{i=1}^n (y_i^* - \hat{y}_i^*)^2 \middle| \mathbf{y} \right) = n^{-1} \sum_{i=1}^n \left(\mathbf{x}_i' (\boldsymbol{\beta} - \mathbf{M} \mathbf{b}_\alpha) \right)^2 + \sigma^2. \quad (3)$$

Note that this error depends on σ^2 , the variability of future observations, and on $(\boldsymbol{\beta} - \mathbf{M} \mathbf{b}_\alpha)$, *i.e.* the error in model selection and estimation of regression coefficients. The cross-validation criterion can be defined as the overall unconditional expected squared error (Shao, 1993):

$$\begin{aligned}\Gamma_{\alpha n} &= \sigma^2 + n^{-1} tr(\mathbf{P}_\alpha) \sigma^2 + n^{-1} \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I}_n - \mathbf{P}_\alpha) \mathbf{X} \boldsymbol{\beta} \\ &= \sigma^2 + n^{-1} \alpha \sigma^2 + n^{-1} \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I}_n - \mathbf{P}_\alpha) \mathbf{X} \boldsymbol{\beta}\end{aligned} \quad (4)$$

where $\mathbf{P}_\alpha = \mathbf{X} \mathbf{M} (\mathbf{M}' \mathbf{X}' \mathbf{X} \mathbf{M})^{-1} \mathbf{M}' \mathbf{X}'$ is the projection matrix of a submatrix $\mathbf{X} \mathbf{M}$. The cross-validation method selects a model by minimizing the estimated $\Gamma_{\alpha n}$ over all the possible subsets of variables, that is all the possible choices of $\mathcal{M}_\alpha$. When a single data set is available, it is split into two parts. The first one, the n_c -size construction sample, is used for fitting a model, whereas the second part, the n_v -size validation sample, is employed

for assessing the predictive ability of the model. The cross-validation technique selects the model with the best average predictive ability calculated on the basis of different combinatory ways of data splitting. The construction of the validation sample is managed by introducing the design matrix $\mathbf{S}$, analogous to matrix $\mathbf{M}$:

$$\mathbf{S} = \left(\mathbf{e}'_{S_1}, \dots, \mathbf{e}'_{S_i}, \dots, \mathbf{e}'_{S_{n_v}} \right)',$$

where $\mathbf{e}_{S_i}$ is the S_i -th column of the identity matrix $\mathbf{I}_n$.

The leave-one-out criterion: CV(1) The leave-one-out criterion, where the validation sample is formed only by one observation, is one of the most widely used. The estimator of (4) is:

$$\begin{aligned} \hat{\Gamma}_{\alpha n}^{CV} &= n^{-1} \sum_{i=1}^n \left((1 - p_{i\alpha})^{-1} \mathbf{S} (\mathbf{y} - \mathbf{X} \mathbf{M} \mathbf{b}_\alpha) \right)^2 \\ &= n^{-1} \sum_{i=1}^n \left((1 - p_{i\alpha})^{-1} \mathbf{S} \mathbf{r}_\alpha \right)^2 \end{aligned} \quad (5)$$

where $\mathbf{S}$ is the matrix selecting the i -th observation as sample validation, $p_{i\alpha}$ is the i -th diagonal element of $\mathbf{P}_\alpha$ and $\mathbf{r}_\alpha$ is the vector of residuals under the model $\mathcal{M}_\alpha$. The only part of (5) related to the selection of the validation sample is the matrix $\mathbf{S}$, while the other components pertain only to the selected model $\mathcal{M}_\alpha$. Shao (1993) shows that the probability of selecting a model which excludes relevant predictors is asymptotically zero. Furthermore, he demonstrates that the probability of including irrelevant predictors is asymptotically different from zero. The same result is obtained by following a different procedure relying on the new notation (for details see Raggi (2002)). In particular, the leave-one-out criterion is asymptotically incorrect in the sense that the probability of selecting the optimal model does not converge to 1 as the total number of observations increases. This method is too conservative: it may select a model of excessive size, unless the optimal model has size p . It is possible to show also that the estimator criterion $\hat{\Gamma}_{\alpha n}^{CV}$ is consistent for $\Gamma_{\alpha n}$ for models in both categories.

The Balanced Incomplete Cross-Validation criterion: BICV(n_v)

The problem of asymptotic bias can be solved by using a validation sample containing more than one observation. The selection of a large dimension validation sample leads quickly to heavy computations, because of the increasing number of possible split samples. Shao (1993) shows that it is unnecessary to carry out validation for every split, but it is possible to adopt an experimental design, like a Balanced Incomplete Block design ($\mathcal{D}$) and to derive the BICV(n_v) (Balanced Incomplete Cross-Validation method). The estimator of criterion (4) is:

$$\hat{\Gamma}_{\alpha n}^{BICV} = \frac{1}{n_v d} \sum_{\mathbf{S}_j \in \mathcal{D}} \left\| \left(\mathbf{I} - \mathbf{S}_j \mathbf{P}_\alpha \mathbf{S}_j' \right)^{-1} \mathbf{S}_j \mathbf{r}_\alpha \right\|^2 \quad (6)$$

where $\mathbf{S}_j$ represents one of the d design blocks. In (5) and (6) it appears that the multiplication by $\mathbf{S}_j$ applies to results which are kept fixed for every selected model, *i.e.* the projection matrix $\mathbf{P}_\alpha$ and $\mathbf{r}_\alpha$. This method is less computationally intensive than extracting all possible validation samples and has the advantage that it is asymptotically unbiased (Shao, 1993).

4. The extension to a wider class of models

The results just described can be extended to the class of Linear Mixed Models (LMM). The uni cation effort performed in the previous section reveals very useful since, within the framework which has been developed above, the extension is straightforward. The Fay and Herriot (1979) model belongs to the LMM class and it is an extension of (1) with: $\varepsilon = (\mathbf{v} + \mathbf{u})$, where $\mathbf{v}_i \stackrel{ind}{\sim} N(0, \mathbf{A})$ are mutually independent from $\mathbf{u}_i \stackrel{ind}{\sim} N(0, \mathbf{D}_i)$. As a consequence $\varepsilon \sim N(0, \Sigma)$ where $\Sigma = \mathbf{A} + \mathbf{D}_i$. For this model the criterion corresponding to (4) is:

$$\Gamma_{\alpha n}^{FH} = tr(\Sigma) + n^{-1}tr(\mathbf{P}_\alpha \Sigma) + n^{-1}\beta' \mathbf{X}' (\mathbf{I}_n - \mathbf{P}_\alpha) \mathbf{X} \beta. \quad (7)$$

Both in (4) and (7) the same last component appears. The differences between criteria depend only on the structure of the error variance-covariance matrix of the two models. The estimator of the cross-validation criterion (7), when a Balanced Incomplete Block design is used, is exactly equal to (6). The calculations show that the different structure of the error variance-covariance matrix lies only in the vector of residuals $\mathbf{r}_\alpha$. Indeed this is an important result which shows the exact point of the solution where the properties typical of the model class contribute to the criterion estimator. The way to provide the asymptotical properties is extended directly to this wider class of models with the following theorem.

Theorem. Under some general conditions (for details see Raggi (2002)) we have:

- for each model $\mathcal{M}_\alpha$ contained in Category I (incorrect model) there exists an $R_n \geq 0$ such that:

$$\hat{\Gamma}_{\alpha n}^{FH} = n^{-1}\varepsilon'\varepsilon + n^{-1}\beta' \mathbf{X}' (\mathbf{I}_n - \mathbf{P}_\alpha) \mathbf{X} \beta + o_p(1) + R_n$$

- for each model $\mathcal{M}_\alpha$ contained in Category II (correct model) the estimated criterion results:

$$\hat{\Gamma}_{\alpha n}^{FH} = n^{-1}\varepsilon'\varepsilon + n_c^{-1}tr(\mathbf{P}_\alpha \Sigma) + o_p(n_c^{-1})$$

- consequently:

$$\lim_{n \rightarrow \infty} P(\mathcal{M}_{FH} = \mathcal{M}_*) = 1$$

where $\mathcal{M}_{FH}$ denotes the cross-validated model. The theorem states that the cross-validation estimator when a Balanced Incomplete Block design is used, is asymptotically unbiased for the Fay and Herriot model. The class of models that can be validated by means of an asymptotically unbiased estimator is therefore extended with this result.

References

- Fay R. and Herriot R. (1979) Estimates of income for small places: an application of james-stein procedures to census data, *Journal of the American Statistical Association*, 74, 269–277.
- Picard R. and Cook R. (1984) Cross-validation of regression models, *Journal of the American Statistical Association*, 79, 575–583.
- Raggi M. (2002) *Asymptotic properties of Cross-Validation methods in regression problems*, PhD thesis. Dipartimento di Scienze Statistiche. Università di Bologna.
- Shao J. (1993) Linear model selection by cross-validation, *Journal of the American Statistical Association*, 88, 486–494.
- Stone M. (1978) Cross-validation: a review, *Math. Operationforsch. Statist., Ser. Statistics*, 9, 127–139.