CROSS-VALIDATION TECHNIQUES IN THE PRACTICAL PROBLEM OF THE CHOICE OF THE REGRESSION ESTIMATOR

Andrzej Grzybowski

Institute of Mathematics and Computer Science, Czestochowa University of Technology

Abstract. The paper is devoted to the problem of a choice between various regression estimators in real world applications. We emphasise the role of cross-validation techniques when doing such a choice in actual usage, especially in the situations where theoretical assumption about considered problem are difficult to verify and the aim of the model building is the prediction of future values of the response variable.

1. Introduction

When one wants to apply the linear regression model to a set of data, there are various methods one could use to estimate the regression coefficients. The most popular one is called the method of least squares (LS). This method, however, has its well-known weaknesses, see e.g. Berger [1], Grzybowski [4]. Consequently there are many situations where we prefer to use alternative regression methods. The decision theory help us here. We are presented with various regression estimators such as Least-Absolute-Deviations regression (LAD), M-(Huber) regression, ridge regression or, incorporating prior information, Bayes, robust Bayes, minimax estimators, to mention at least the most popular ones - see e.g. [1,4]. However, the optimal performance of the estimators depends on the problem formulation and various assumptions about the model. In actual usage it is often difficult to decide what description of the problem is most appropriate, and consequently, which estimator is most suitable. In this paper we focus on such situations and emphasise the role of cross-validation simulations in choosing the estimator when the prediction is the main purpose of the model building.

2. Problem formulation

Considered models have the usual linear form: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}$, where $\mathbf{Y}$ is an n-dimensional vector of observations of the dependent variable, $\mathbf{X}$ is a given non-stochastic (n x k) matrix with the rank k, β is a k-dimensional vector of unknown regression coefficients, $\mathbf{Z}$ is an n-dimensional vector of random disturbances.

To illustrate the usage of the cross validation techniques for the choice of the estimator we examine two data sets. One is connected with the model of monthly rental price built on the base of measurements made for 36 apartments and is taken from the book Frees [3]. The second set of data was gathered during the forming and rolling process in Częstochowa steel works. The data set was obtained for the conventional method of rolling and consists of 50 records.

In our paper we use only *two* estimators LS and LAD to present the idea of the actual data based choice of the method of estimation. However, it is obvious that it may be applied to an arbitrary number of estimators. The formulae for the estimators may be found e.g. in Birkes, Dodge [2].

3. Cross-validation techniques

This section provides an introduction to a variety of CV heuristics. In the sequel ”to select a model” means ”to select an estimator which lead to a given estimates of the model coefficients”.

Hold out set CV. ”Hold out set” CV (hereafter referred to as HOS-CV) is perhaps the most obvious form of CV. The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the estimation (or training) set and the validation (or testing) set. The estimator fits a function using the estimation set only. Then on the base of the obtained model the output values are predicted for the data in the validation set. The errors it makes are accumulated as before to give the mean absolute test set error, which is used to evaluate the model. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data records end up in the training set and which end up in the validation set, and thus the evaluation may be significantly different depending on how the division is made.

Multifold CV. HOS-CV is often sufficient when abundant data are available. However, when it is not known whether or not the available amount of data is sufficient to ensure proper training and still provide an adequate amount of data for testing, it is well known that HOS-CV can be sensitive to user-set parameters that determine the size of the validation set (and hence, as well, the size of the training set) as well as the choice of split (i.e., a particular distribution of the available records into either the training set or the validation set). Let N_t and $N_v = N\text{-}N_t$ denote the sizes of the training set and validation set, respectively. Given choice of N_v and N_t, one direct method to reduce the sensitivity of HOS-CV on the choice of split is to repeat the HOS-CV procedure a number of times and average the results. We refer to this method generically as ”multifold” CV. Multifold CV can be implemented in several ways. One way is ”Leave-Many-Out” CV (LMO-CV) Note that there are $C(N, N_v)$ ways to select a validation set of size N_v, where

$$C(k,n) = k!/[k!(n\text{-}k)!] \qquad (1)$$

gives the number of combinations available by choosing *k* things out of *n* that are available.

Here are three methods of implementing LMO-CV:

1) ”Exhaustive” LMO-CV (ELMO-CV): average over all $C(N, N_v)$ splits.
2) ”Disjoint set” LMO-CV (DLMO-CV): split the data into $[N/N_v]$ disjoint validation subsets of size N_v i.e., take the ”floor” of N/N_v, rounding it down to the nearest integer.
3) ”Monte-Carlo” LMO-CV (MCLMO-CV): select each validation subset by drawing N_v records at random without replacement (cf. [Shao 93]).

ELMO-CV can be prohibitively expensive for even modest amounts of data, as $C(N,N_v)$ grows large quickly with N for $N_v>1$. DLMO-CV is the implementation of multifold CV very commonly used in literature (for instance, it is essentially equivalent to an implementation known as ”V-fold CV,” where V = $[N/N_v]$ gives the number of disjoint validation sets, or, ”folds”). MCLMO-CV was suggested by Shao [7] and has some appealing (and possibly surprising) properties. MCLMO-CV is similar to the bootstrap in that a particular estimation set could be generated more than once. However, note that MCLMO-CV differs from the conventional bootstrap, where training sets are obtained by sampling from the available data at random *with* replacement. Both ELMO-CV and MCLMO-CV average over many more validation sets than DLMO-CV, in general, because they generate many more splits than does DLMO-CV.

Leave-One Out CV. All of the multifold methods involve choice of N_v. Setting $N_v=1$ yields the popular and well-studied “Leave-One-Out” CV (LOO-CV). LOO--CV may be considered as a special case of LMO-CV. Perhaps LOO is an appealing alternative to LMO-CV to many practitioners because:

- it eliminates the need to select the value of *M*,
- it limits the number of splits (to a total of *N*), and
- it results in the maximum number of records (i.e., *N*-1) allotted to each training set.

However, these reasons can be outweighed by the much better performance that can result by using LMO-CV, as is discussed e.g. in Plutowski, Sakata, White [6].

While CV is generally applicable, it was once considered too expensive to apply directly in many practical settings (although this is less of a concern now that computation is becoming less expensive). Although LOO-CV limits the number of training subsets to *N*, the cost of fitting the model to each of the *N* sets may still be prohibitively expensive. Other properties of the considered techniques one may find in Kohavi [5].

In our examples we make use of the two most effective cross-validation methods: LOO-CV and MCLMO-CV.

4. Results

We use now the above described techniques LOO-CV and MCLMO-CV to choose better from the LS and LAD estimators in two problems of regression estimation. First one is connected with model of the steel plates properties, the second deals with monthly apartment rents. It appeared that in both cases the standard assumptions about linear regression models cannot be rejected and the LS estimator may be used. However let us assume that we are not sure what form of the loss is appropriate. Thus the comparison is made for two loss functions: the Euclidian norm of the difference between the parameter β end its estimate **b** given as

$$L_1(\beta,\mathbf{b}) = \sqrt{(\mathbf{b}-\beta)^T(\mathbf{b}-\beta)} \qquad (2)$$

and for the quadratic loss function given by

$$L_2(\beta,\mathbf{b}) = (\boldsymbol{b}-\beta)^T(\boldsymbol{b}-\beta) \qquad (3)$$

Problem 1

The aim of the example is to compare the aforementioned methods as tools for estimation the regression parameter in models for mechanical properties of steel plates. We want to model so called the *yield stress Re* as a function of some technological properties of the rolling process and the chemical composition of the plate.

After preliminary studies we set up significant explanatory variables and assume the following form of the model:

$$Re = \beta_1 + \beta_2 Mn + \beta_3 Si + \beta_4 Ni + \beta_5 Al + \beta_6 Th + \beta_7 ERT + Z \qquad (4)$$

In the above formula Z stands for random disturbance, symbols *ERT* and *Th* stand for the *end of rolling temperature* and *thickness*, respectively. The remaining are chemical symbols. Our analysis is based on the data gathered in real terms during the forming and conventional rolling process. The data set consists of 50 records which can be found in Grzybowski [4].

To compare the LS and LAD estimators of the parameter $\beta = (\beta_1,\dots,\beta_7)^T$ we first use the LOO-CV technique. So the presented results are an average loss obtained for 50 problems. Next the MCLMO-CV method with $N_v = 20$ is adopted. The random split of the data set was generated a hundred times. Table 1 shows the result obtained in this case. Apart from the mean values of loss functions L_2 and L_1 we also compute the value of the statistic

$$U = \frac{\overline{L}_i^{LAD} - \overline{L}_i^{LS}}{\sqrt{S_i^2(LAD) + S_i^2(LS)}}\sqrt{n} \qquad (5)$$

to test the hypothesis that the mean values of loss (L_1 or L_2) for given estimators are actually equal. Here the symbols L_i^{LAD}, L_i^{LS} and $S_i^2(LAD)$, $S_i^2(LS)$ stand for the loss L_i, $i = 1,2$, and its standard deviation, computed for estimator *LAD* or *LS*, respectively, *n* is a number of predictions made during the simulations. The p-value presented in the table is an observed significance level - the probability, assuming the null hypothesis is true, of observing a value of the test statistic that is at least as contradictory to the null hypothesis and as supportive of the alternative hypothesis, as the one we have computed.

Table 1. The comparison of LS and LAD estimators for the steel plate properties model

CV method	*LS* estimator		*LAD* estimator		U statistic		p-value	
	L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2
LOO-CV	22.18	921.99	19.38	687.97	–3.198	–27.24	0.0007	0
MCLMO-CV	25.4	1473.6	19.94	751.2	–35.39	–401.3	0	0

We see that both methods of cross validation show that the LAD estimator is significantly better for estimating regression parameter for this model - no matter what the loss function is under consideration.

Problem 2

In this example we analyse data connected with monthly apartment rental prices downtown Medison, USA. The data may be found in the book of Frees [3]. There the following model for *apartment monthly rent per square meter* (*RpM*) is proposed:

$$RpM = \beta_1 + \beta_2 Dis + \beta_3 Size + \beta_4 TB + Z \qquad (6)$$

where *Dis* is the distance from the city centre, in kilometres, *Size* is an apartment size measured in square feet and, finally, *TB* is the type of apartment: 1 if a two bedroom and 0 if a one bedroom, *Z* is the random disturbance. Table 2 presents the results obtained for the model.

Table 2. The comparison of LS and LAD estimators for the model of monthly rental prices

CV method	*LS* estimator		*LAD* estimator		U statistic		p-value	
	L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2
LOO-CV	0.92	1.25	1.12	1.65	1.06	1.26	0.14	0.1
MCLMO-CV	0.93	1.3	0.98	1.41	1.43	1.73	0.076	0.042

This time both CV methods indicate the better performance of the model based on LS estimates. Thus the estimator should be suggested for use, at least in the case where the prediction is the aim of model building.

5. Final remarks

In the paper we present an example of the practical usage of the cross validation techniques in order to choose regression estimator which is most suitable for a given data set. However we should stress that the criteria of performance of the compared estimators are based on the prediction error and thus if the principal aim of the estimation is not the prediction, the results may be misleading. Such a situation occurs e.g. when we want to investigate the relationship between response and some of explanatory variables rather than to predict future values of the response. It is well known from decision theory that the prediction and estimation task are usually not exchangeable in that sense, that the estimators better for one task does not have to be better for another, see e.g. Grzybowski [4].

References

[1] Berger J.O., Statistical Decision Theory and Bayesian Analysis, Springer Verlag, New York 1985.

[2] Birkes D., Dodge Y., Alternative methods of regression, Wiley & Sons, New York 1993.

[3] Frees E.W., Data analysis using regression models - the business perspective, Prentice-Hall Inc., New Jersey 1996.

[4] Grzybowski A., Metody wykorzystania informacji a priori w estymacji parametrów regresji, Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2002, seria monografie.

[5] Kohavi R., A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Mateo 1995, 1137-1143.

[6] Plutowski M.E., Sakata S., White H., Cross-validation estimates integrated mean squared error, Giles C.L., Hanson S.J., Cowan J.D. (eds.), Advances in neural information processing systems 6, Morgan Kaufmann Publishers, San Mateo 1994.

[7] Shao, J., Linear model selection by cross-validation, Journal of the American Statistical Association 1993, 88, 422, 486-494.