Lasso-based linear regression for interval-valued data

Giordani, Paolo
Sapienza University of Rome, Department of Statistical Sciences
P.le Aldo Moro, 5
00185 Rome, Italy
E-mail: paolo.giordani@uniroma1.it

Abstract

In regression analysis the relationship between one response and a set of explanatory variables is investigated. The (response and explanatory) variables are usually single-valued. However, in several real-life situations, the available information may be formalized in terms of intervals. An interval-valued datum can be described by the midpoint (its center) and the radius (its half width). Here, limiting our attention to the linear case, regression analysis for interval-valued data is studied. This is done by considering two linear regression models. One model investigates the relationship between the midpoints of the response variable and of the explanatory variables, whereas the other one analyzes the relationship between the radii. The two models are related by considering the same regression coefficients, i.e. the same linear relationship is assumed for the midpoints and the radii. However, in some cases, this assumption may be too restrictive. To overcome this drawback, additive coefficients for the model of the radii are introduced and their magnitude is tuned according to the Lasso technique allowing us to set to zero some of these additive coefficients. In order to show how the proposed method works in practice the results of an application to real-life data are discussed.

Keywords

Interval-valued data, Linear regression analysis, Lasso technique.

Introduction

In the presence of interval-valued data, the attributes involved can be expressed by a lower and an upper bound, $\underline{z}$ and $\overline{z}$ with $\overline{z} \geq \underline{z}$, respectively, providing the boundaries of the interval-valued data. However, interval-valued data are usually expressed in terms of the so-called midpoint and radius, say z_M and z_R (≥ 0), with $z_M = \frac{(\overline{z}+\underline{z})}{2}$ and $z_R = \frac{(\overline{z}-\underline{z})}{2}$. The midpoint is the center of an interval (the location), whereas the radius is the half-width of an interval (a measure of the imprecision). In this work, the linear regression problem for interval-valued data is investigated. In the literature, the topic has been deeply analyzed. See, for instance, González-Rodríguez et al. (2007), Blanco et al. (2008, 2010), Lima-Neto and De Carvalho (2008, 2010). The peculiarity of the here-proposed model is that the Lasso technique (Tibshirani, 1996) is considered in order to get regression coefficients for the midpoints close as much as possible to the corresponding ones for radii as it will be clarified in the next section that contains the details on the regression model. Then, an algorithm to estimate the parameters of the regression model is provided. Finally, the results of an application are discussed and some concluding remarks are given.

Model

Let Y and $X_1, \ldots, X_p$ be an interval-valued response variable and a number of interval-valued explanatory variables, respectively, observed on n units. The linear relationship between Y and

$X_1, \ldots, X_p$ can be written as

$$(1) \qquad \begin{array}{l} \mathbf{y}_M = \mathbf{y}_M^* + \mathbf{e}_M = \mathbf{X}_M \mathbf{b}_M + \mathbf{e}_M \text{ (midpoint model)}, \\ \mathbf{y}_R = \mathbf{y}_R^* + \mathbf{e}_R = \mathbf{X}_R \mathbf{b}_R + \mathbf{e}_R = \mathbf{X}_R \left(\mathbf{b}_M + \mathbf{b}_A\right) + \mathbf{e}_R \text{ (radius model)}, \end{array}$$

where $\mathbf{y}_M$ and $\mathbf{y}_R$ denote the vectors of length n of the observed midpoints and of the observed radii of the response variable, respectively, and $\mathbf{y}_M^*$ and $\mathbf{y}_R^*$ are the vectors of the theoretical midpoints and radii of the response variable, respectively. $\mathbf{X}_M$ and $\mathbf{X}_R$ are, respectively, the matrices of order $(n \times p + 1)$ of the midpoints and of the radii of the explanatory variables containing the unit vector of length n in their first column. $\mathbf{e}_M$ and $\mathbf{e}_R$ denote the residual vectors. Finally, $\mathbf{b}_M$ and $\mathbf{b}_R$ are the vectors of length $(p+1)$ of the regression coefficients for the midpoint and radius models, respectively, where $\mathbf{b}_R = \mathbf{b}_M + \mathbf{b}_A$ being $\mathbf{b}_A$ the vector of the additive coefficients. In (1) the coefficients of the radius model $\mathbf{b}_R$ are equal to the corresponding ones of the midpoint model $\mathbf{b}_M$ up to the additive coefficients $\mathbf{b}_A$. Such a model takes inspiration from the idea underlying González-Rodríguez et al. (2007) in the sense that the attempt to seek a common set of regression coefficients for the midpoint and the radius model is pursued even if *to some extent*. This is done by adding specific regression coefficients that, however, are constrained to be as small as possible according to a tuning parameter to be chosen by the researcher.

The parameter vectors $\mathbf{b}_M$ and $\mathbf{b}_A$ are estimated in such a way to minimize a suitable dissimilarity measure between observed and theoretical data. For this purpose, the squared distance d_θ^2 proposed by Trutschnig et al. (2009) is considered. Given two intervals $G \equiv (G_M, G_R)$ and $H \equiv (H_M, H_R)$ it is

$$(2) \qquad d_\theta^2 = \left(G_M - H_M\right)^2 + \theta \left(G_R - H_R\right)^2$$

with $\theta \in (0, 1]$. When $\theta = 1$, d_θ^2 compares G and H by the sum of the squared distances of their midpoints and of their radii. The choice of θ depends on the relative importance of the radius distance with respect to the midpoint distance. A reasonable choice seems to be $\theta = \frac{1}{3}$.

Using (2), the loss function to be minimized is

$$(3) \qquad \min_{\mathbf{b}_M, \mathbf{b}_A} \left\|\mathbf{e}_M\right\|^2 + \theta \left\|\mathbf{e}_R\right\|^2 = \left\|\mathbf{y}_M - \mathbf{X}_M \mathbf{b}_M\right\|^2 + \theta \left\|\mathbf{y}_R - \mathbf{X}_R \left(\mathbf{b}_M + \mathbf{b}_A\right)\right\|^2.$$

The loss function in (3) requires some constraints in order to guarantee that the estimated radii are non-negative and that the additive coefficients $\mathbf{b}_A$ are as small as possible. The former requirement can be achieved setting

$$(4) \qquad \mathbf{X}_R \left(\mathbf{b}_M + \mathbf{b}_A\right) \geq \mathbf{0}.$$

The latter requirement can be managed using the Lasso technique (Least Absolute Shrinkage and Selection Operator) proposed by Tibshirani (1996). Lasso is a well-known method used in standard regression analysis aiming at shrinking some regression coefficients and setting some others to 0. This is done by minimizing the residual sum of squares with the constraint that the sum of the absolute values of the regression coefficients is smaller than a threshold. It can be shown that the minimization problem to be solved by Lasso is a quadratic programming problem with linear inequality constraints, the solution of which can be found in Lawson and Hanson (1995). The use of the Lasso constraint in the here-proposed model for interval-valued data can be carried out introducing the following constraint:

$$(5) \qquad \sum_{j=0}^{p} \left|b_{Aj}\right| \leq t.$$

This allows us to limit the magnitude of the additive coefficients as much as possible according to the choice of t. Note that, differently from the standard Lasso, in (5) the Lasso constraint is considered for all the (additive) coefficients including the intercept. Taking into account (3), (4) and (5) we then get the following constrained minimization problem:

$$\begin{aligned}&\min_{\mathbf{b}_M,\mathbf{b}_A} \|\mathbf{y}_M - \mathbf{X}_M\mathbf{b}_M\|^2 + \theta\,\|\mathbf{y}_R - \mathbf{X}_R(\mathbf{b}_M + \mathbf{b}_A)\|^2,\\ &\text{s.t. } \mathbf{X}_R(\mathbf{b}_M + \mathbf{b}_A) \geq \mathbf{0},\ \sum_{j=0}^{p} |b_{Aj}| \leq t.\end{aligned} \tag{6}$$

We refer to (6) as Lasso-based Interval-valued Regression (Lasso-IR).

Before minimizing (6), the shrinkage parameter t must be fixed. The possible values of t range from 0 to t_{MAX}. When $t = 0$ it is $\mathbf{b}_A = \mathbf{0}$ and, therefore, the same regression coefficients for the midpoints and the radii are found. t_{MAX} is the smallest value such that two *separate* regression problems for the midpoint and the radius models are solved. Of course, if $t > t_{\text{MAX}}$ is chosen the same solution for the case with $t = t_{\text{MAX}}$ is obtained. The value of t can be chosen either on the basis of the experience of the researcher or according to cross-validation techniques, such as the k-fold cross-validation procedure (see, for instance, Efron and Tibshirani, 1993). In Lasso-IR for different values of t ranging from 0 to t_{MAX} we can compute the predictive accuracy as

$$CV(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_{Mi} - \widehat{y_{Mi}}^{(-k(i))}(t)\right)^2 + \theta\left(y_{Ri} - \widehat{y_{Ri}}^{(-k(i))}(t)\right)^2\right], \tag{7}$$

where $\widehat{y_{M}}_i^{(-k(i))}$ and $\widehat{y_{R}}_i^{(-k(i))}$ denote the i-th fitted midpoint and radius, respectively, computed setting t and removing the k-th part of the data. Then the optimal value of t is

$$t_{\text{OPT}} = \underset{0\leq t\leq t_{\text{MAX}}}{\arg\min}\ CV(t). \tag{8}$$

See, for more details about LASSO-IR, Giordani (2011).

Algorithm

To solve the minimization problem in (6) an alternating least squares algorithm is proposed. First, initial values for $\mathbf{b}_A$ fulfilling the constraints in (6) must be given. For instance, these can be found randomly from U[0,1] rescaling them if necessary. Then, the algorithm consists of updating separately the vectors $\mathbf{b}_M$ and $\mathbf{b}_A$ keeping fixed the remaining one. Whenever a vector is updated, the loss function to be minimized decreases. After updating both the vectors, if the loss function value decreases less than a specified percentage (e.g. 0.0001%) from the previous function value, we consider the algorithm converged, otherwise we repeat the updates of $\mathbf{b}_M$ and $\mathbf{b}_A$. The function in (6) has a lower bound and, therefore, the function value converges to a stable value.

The updates of $\mathbf{b}_M$ can be found noting that the constraints in (6) do not play an active role in the update of $\mathbf{b}_M$. After a little algebra (6) can be rewritten as

$$\left\|\begin{bmatrix}\mathbf{y}_M\\ \theta^{1/2}(\mathbf{y}_R - \mathbf{X}_R\mathbf{b}_A)\end{bmatrix} - \begin{bmatrix}\mathbf{X}_M\\ \theta^{1/2}\mathbf{X}_R\end{bmatrix}\mathbf{b}_M\right\|^2 = \|\mathbf{c} - \mathbf{D}\mathbf{b}_M\|^2, \tag{9}$$

where $\mathbf{c}$ and $\mathbf{D}$ are implicitly defined in (9), from which we get

$$\widehat{\mathbf{b}_M} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{c}. \tag{10}$$

In order to update $\mathbf{b}_A$ and starting from (6) it is easy to see that the problem to be solved reduces to

$$\begin{aligned}&\min_{\mathbf{b}_A} \|\mathbf{y}_R - \mathbf{X}_R(\mathbf{b}_M + \mathbf{b}_A)\|^2,\\ &\text{s.t. } \mathbf{X}_R(\mathbf{b}_M + \mathbf{b}_A) \geq \mathbf{0},\ \sum_{j=0}^{p} |b_{Aj}| \leq t,\end{aligned} \tag{11}$$

keeping $\mathbf{b}_M$ fixed. The problem in (11) can be recognized as a constrained regression problem where the response variable is $\mathbf{y}_R - \mathbf{X}_R\mathbf{b}_M$ and the explanatory ones are $\mathbf{X}_R$. It can be shown that (11) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{b}_A} \left\| (\mathbf{y}_R - \mathbf{X}_R\mathbf{b}_M) - \mathbf{X}_R\mathbf{b}_A \right\|^2, \\ &\text{s.t.} \begin{bmatrix} \mathbf{X}_R \\ \mathbf{H} \end{bmatrix} \mathbf{b}_A \geq \begin{bmatrix} -\mathbf{X}_R\mathbf{b}_M \\ t\mathbf{1}_{2^{p+1}} \end{bmatrix}, \end{aligned} \tag{12}$$

in which $\mathbf{1}_{2^{p+1}}$ is the unit column vector of length 2^{p+1} and $\mathbf{H}$ is a $(2^{p+1} \times p+1)$ matrix containing in its rows all the 2^{p+1} combinations of length $(p+1)$ of ± 1. In the literature there exist several methods to solve (12). See, for instance, Lawson and Hanson (1995) and Gill et al. (1981). For further details about the iterative algorithm see Giordani (2011).

Application

In this section the results of an application to real data are discussed. The data set refers to the values of three cardiological variables, namely the pulse rate, the systolic pressure and the diastolic pressure observed on a set of patients. In the literature, the data can be found in Billard and Diday (2000) and, for the convenience of the reader, is reported in the following table.

Cardiological data set (Billard and Diday, 2000)

Patient	Pulse rate	Systolic pressure	Diastolic pressure
1	[44,68]	[90,100]	[50,70]
2	[60,72]	[90,130]	[70,90]
3	[56,90]	[140,180]	[90,100]
4	[70,112]	[110,142]	[80,108]
5	[54,72]	[90,100]	[50,70]
6	[70,100]	[130,160]	[80,110]
7	[63,75]	[140,150]	[60,100]
8	[72,100]	[130,160]	[76,90]
9	[76,98]	[110,190]	[70,110]
10	[86,96]	[138,180]	[90,110]
11	[86,100]	[110,150]	[78,100]

The recorded values take the form of intervals and concern $n = 11$ patients. In order to study the linear dependence of the pulse rate (Y) with respect to the systolic pressure (X_1) and the diastolic pressure (X_2) we applied Lasso-IR. Since the number of units is low, the leave-one-out procedure (i.e. k-fold with $k = 1$) has been considered for determining the tuning parameter t, where t ranged from 0 to $t_{\mathrm{MAX}} = 1.29$. with increasing step equal to 0.01. The minimum value of $CV(t)$ was obtained when $t_{\mathrm{OPT}} = 0.79$. By setting $t = 0.79$ we got $\widehat{b_M} = \begin{pmatrix} 11.12 & -0.07 & 0.90 \end{pmatrix}'$ and $\widehat{b_A} = \begin{pmatrix} 0 & 0 & -0.79 \end{pmatrix}'$ from which

$$\begin{aligned} Y_M &= 11.12 - 0.07X_{1M} + 0.90X_{2M}, \\ Y_R &= 11.12 - 0.07X_{1R} + 0.11X_{2R}. \end{aligned}$$

The value of $\widehat{b_{A2}}$ suggested that there exist different linear relationships between Y and X_2 for the midpoints and for the radii. Specifically, a positive relationship was found for the medium levels (i.e. the midpoints of the intervals) of the pulse rate and the diastolic pressure ($\widehat{b_{M2}} = 0.90$). Such a strong relationship did not hold for the variations (i.e. the radii). In fact, the corresponding estimated regression coefficient was $\widehat{b_{R2}} = 0.11$. Conversely, since $\widehat{b_{A1}} = 0$ the same relationship for the midpoints and the radii was found with regard to Y and X_1. In particular, a negative relationship was obtained between pulse rate and systolic pressure ($\widehat{b_{M1}} = \widehat{b_{R1}} = -0.07$). Finally, no distinction for the intercepts of the midpoint and radius models was found.

Final remarks

In this paper we proposed a tool called Lasso-IR for performing linear regression analysis of interval-valued data. It consists of two regression models, one for the midpoints of the intervals and the other one for the radii. The two regression models are characterized by the same regression coefficients *as much as possible* according to a given criterion based on the Lasso technique. A unique set of coefficients is desirable for the sake of parsimony. Unfortunately, this can limit the applicability of the model in some cases. Thus, to make the model more flexible, the regression coefficients for the radii are allowed to differ to some extent from the corresponding ones for the midpoints. This is achieved by introducing additive coefficients for the radius model such that their sum in absolute value is not bigger than a shrinking parameter t that can be either fixed in advance or chosen by cross-validation techniques.

REFERENCES (RÉFERENCES)

References

[Billard and Diday (2000)] Billard, L., Diday, E., 2000. Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.-P., Groenen, P.J.F., Schader, M., (Eds.), Data Analysis, Classification and Related Methods. Springer-Verlag, Heidelberg, pp. 369–374.

[Blanco et al. (2008)] Blanco, A., Corral, N., Colubi, A., González-Rodríguez, G., 2008. On a linear independence test for interval-valued random sets. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.A., Grzegorzewski, P., Hryniewics, O., (Eds.), Soft Methods for Handling Variability and Imprecision. Springer-Verlag, Heidelberg, pp. 331–337.

[Blanco et al. (2010)] Blanco, A., Corral, N., González-Rodríguez, G., Palacio, A., 2010. On some confidence regions to estimate a linear regression model for interval data. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano M.A., Gil, M.A., Grzegorzewski, P., Hryniewicz, O., (Eds.), Combining Soft Computing and Statistical Methods in Data Analysis. Springer-Verlag, Heidelberg, 263–271.

[Efron and Tibshirani (1993)] Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Chapman & Hall, New York.

[Gill et al. (1981)] Gill, P.E., Murray, W., Wright, M.H., 1981. Practical Optimization. Academic Press, London.

[Giordani (2011)] Giordani, P., 2011. Linear regression analysis for interval-valued data based on the Lasso technique. Technical Report n. 7, Department of Statistical Sciences, Sapienza University of Rome.

[González-Rodríguez et al. (2007)] González-Rodríguez, G., Blanco, A., Corral, N., Colubi, A., 2007. Least squares estimation of linear regression models for convex compact random sets. Advances in Data Analysis and Classification 1, 67–81.

[Lawson and Hanson (1995)] Lawson, C.L, Hanson, R.J., 1995. Solving Least Squares Problems, (Classics in Applied Mathematics, Vol. 15). SIAM, Philadelphia.

[Lima-Neto and De Carvalho (2008)] Lima-Neto, E.A., De Carvalho, F.A.T., 2008. Centre and range method to fitting a linear regression model on symbolic interval data. Computational Statistics and Data Analysis 52, 1500–1515.

[Lima-Neto and De Carvalho (2010)] Lima-Neto, E.A., De Carvalho, F.A.T., 2010. Constrained linear regression models for symbolic interval-valued variables. Computational Statistics and Data Analysis 54, 333–347.

[Tibshirani (1996)] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society - Series B 58, 267–288.

[Trutschnig et al. (2009)] Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A., 2009. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences, 179 3964–3972.