Simultaneous confidence regions in a flexible linear regression for random intervals: a bootstrap approach.

Blanco-Fernández, Angela
University of Oviedo, Statistics and Operational Research Department
C/ Calvo Sotelo s/n
Oviedo (33007), Asturias, Spain
E-mail: blancoangela@uniovi.es

Colubi, Ana
University of Oviedo, Statistics and Operational Research Department
C/ Calvo Sotelo s/n
Oviedo (33007), Asturias, Spain
E-mail: colubi@uniovi.es

González-Rodríguez, Gil
University of Oviedo, Statistics and Operational Research Department
C/ Calvo Sotelo s/n
Oviedo (33007), Asturias, Spain
E-mail: gil@uniovi.es

1. Introduction

When real experimental data take interval values (representing, for instance, fluctuations or variations of a magnitude along a given period of time), they are usually modeled by means of the so-called interval-valued random variables, or random intervals.

The linear relationship between two random intervals can be formalized through an interval arithmetic-based linear model with flexible and versatile properties. The LS estimation of this model has been addressed in Blanco-Fernandez et al., 2011. The closed form LS estimators are found as the solution of an optimization problem with constraints.

Once the estimation process of the linear model has been solved, the development of other inferential studies about the model can be addressed. In the interval scenario exact parametric methods are not feasible yet for inferential studies, since no realistic parametric models to describe the distribution of the random intervals have been shown to be widely applicable in practice. Inferential studies for the linear model can be developed by means of asymptotic techniques, based on the study of the limit distributions of the regression estimators (see, for instance, Gil et al., 2007). To improve the results for finite sample sizes, bootstrap methods are widely considered (see, for instance, Colubi, 2009, Blanco-Fernandez et al., 2010).

In this work, a bootstrap algorithm for the construction of simultaneous confidence regions for the regression function is proposed. The procedure is based on the classical method of paired bootstrap, since both intervals in the model are considered as random elements (see Efron & Tibshirani, 1993).

The rest of the paper is organized as follows: in Section 2 some preliminary concepts in the interval scenario are introduced. Then, the formalization and the LS estimation of the simple linear model between random intervals is presented in Section 3. The construction of simultaneous confidence regions for the regression function of the model is shown in Section 4, and a bootstrap algorithm for the practical computation of those confidence regions is proposed. The empirical performance of the suggested algorithm is illustrated in Section 5, by means of some simulation studies and its application over a real-life example. Finally, Section 6 collects some conclusions and future directions.

2. Preliminaries

Let $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ be the space of nonempty compact intervals of $\mathbb{R}$ endowed with the semilinear structure induced by the Minkowski addition and the product by a scalar, that is, $A+B = \{a+b \,|\, a \in A\,, b \in B\}$ and $\lambda A = \{\lambda a \,|\, a \in A\}$ for all $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. Given $A, B \in \mathcal{K}_c(\mathbb{R})$, if there exists $C \in \mathcal{K}_c(\mathbb{R})$ so that $A = B + C$, then C is defined as the Hukuhara difference between A and B, denoted by $A -_H B$. Each interval $A \in \mathcal{K}_c(\mathbb{R})$ can be characterized by means of the real vector $(\inf A, \sup A) \in \mathbb{R}^2$ such that $\inf A \leq \sup A$. Equivalently, A can be parametrized through its midpoint (or centre) $\text{mid}A$ and its spread (or radius) $\text{spr}A$, by means of the real vector $(\text{mid}A, \text{spr}A)$ with $\text{spr}A \geq 0$, where $\text{mid}A = (\sup A + \inf A)/2$ and $\text{spr}A = (\sup A - \inf A)/2$. The notation $A = [\inf A, \sup A]$ or $A = [\text{mid}A \pm \text{spr}A]$, respectively, will be considered in each case.

Several metrics can be defined on the space $\mathcal{K}_c(\mathbb{R})$. For least squares problems associated with regression studies, an L_2-type metric is suitable. A generalized L_2-type distance between two intervals A and B can be defined as

$$(1) \quad d_\theta(A, B) = \sqrt{(\text{mid}A - \text{mid}B)^2 + \theta(\text{spr}A - \text{spr}B)^2}$$

for an arbitrary $\theta > 0$ (see Trutschnig et al., 2009). The constant θ allows us to choose the relative importance given for the squared Euclidean distance between the spreads of the intervals (which can be seen as a measure of the difference in imprecision of both intervals) with respect to the squared Euclidean distance between the midpoints (which measures the difference in location or position in the real line) of the intervals. In particular, $d_{1/3}$ is often employed for applications, since it corresponds to the Bertoluzza metric d_W (see Bertoluzza et al., 1995), when W is the Lebesgue measure on $[0, 1]$.

Given a probability space $(\Omega, \mathcal{A}, P)$, a mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ is said to be an *interval-valued random set* (or *random interval*), if it is $\mathcal{A}|\mathcal{B}_{d_\theta}$-measurable, $\mathcal{B}_{d_\theta}$ denoting the σ-field generated by the topology induced by the metric d_θ on $\mathcal{K}_c(\mathbb{R})$.

Let $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ be a random interval such that $E(|X|) < \infty$ (with $|X|(\omega) = \sup\big\{|x| \,\big|\, x \in X(\omega)\big\}$ for all $\omega \in \Omega$), then, the *expected value of* X *in Kudō-Aumann's sense* (see Aumann, 1965) is the interval $E(X) = [E(\text{mid}X) \pm E(\text{spr}X)]$. The considered second-order moments for random intervals are the usual (Fréchet) ones in metric spaces. Specifically, in terms of the mid and spread variables of the random intervals X and Y the variance and the covariance can be written as $\sigma^2_X = \sigma^2_{\text{mid}\,X} + \theta\sigma^2_{\text{spr}\,X}$ and $\sigma_{X,Y} = \sigma_{\text{mid}\,X,\text{mid}\,X} + \theta\sigma_{\text{spr}\,X,\text{spr}\,X}$, respectively, whenever the corresponding real moments exist. For $\{X_i, Y_i\}_{i=1}^n$ a random sample obtained from (X, Y), the sample moments can be defined in the usual way, that is: $\overline{X} = (X_1 + \ldots + X_n)/n$, $\widehat{\sigma}^2_X = (\sum_{i=1}^n d^2_\theta(X_i, \overline{X}))/n = \widehat{\sigma}^2_{\text{mid}\,X} + \theta\,\widehat{\sigma}^2_{\text{spr}\,X}$ (analogously for Y) and $\widehat{\sigma}_{X,Y} = \widehat{\sigma}_{\text{mid}\,X,\text{mid}\,X} + \theta\,\widehat{\sigma}_{\text{spr}\,X,\text{spr}\,X}$.

3. A flexible linear regression model for random intervals

The formalization of the linear model for random intervals presented in Blanco-Fernández et al., 2011 is based on the *canonical decomposition* of the intervals. The notation $A = [\text{mid}A \pm \text{spr}A]$ can be split into two terms depending on the midpoint and spread values of A as $A = \text{mid}\,A[1 \pm 0] + \text{spr}\,A[0 \pm 1]$. This expression allows us to work separately with the *mid* and *spr* components of the interval, but keeping the interval arithmetic.

Let X and Y be two random intervals with finite second-order moments, and $\text{spr}\,X$ non-degenerated (so X is not reduced to a real-valued random variable). Based on the canonical decomposition, the linear model between X and Y is formalized as

$$(2) \quad Y = \alpha\text{mid}X[1 \pm 0] + \beta\text{spr}X[0 \pm 1] + \gamma[1 \pm 0] + \varepsilon\ ,$$

where α and β are the regression coefficients, γ is an intercept term affecting the *mid* component of Y, and ε is an interval-valued random error variable such that $E(\varepsilon|X) = [-\delta, \delta] \in \mathcal{K}_c(\mathbb{R})$ (so that $\delta \geq 0$).

For simpler notation, if we define $B = [\gamma - \delta, \gamma + \delta] \in \mathcal{K}_c(\mathbb{R})$, the regression function associated with the model (2) will be denoted by

$$(3) \quad E(Y|X) = \alpha X^M + \beta X^S + B,$$

where $X^M = \operatorname{mid} X[1 \pm 0]$ and $X^S = \operatorname{spr} X[0 \pm 1]$. Since the random interval X^S verifies that $X^S = -X^S$, it is possible to assume without loss of generality that $\beta \geq 0$.

The LS estimation of the model (2) has been solved by means of a minimization problem over a suitable feasible set assuring the existence of the residuals of the sample model. Given a random sample $\{X_i, Y_i\}_{i=1}^n$ from (X, Y), the following analytic expressions for the estimators for the parameters α, β and B of the regression function (3) are obtained:

$$(4) \quad \widehat{\alpha} = \frac{\widehat{\sigma}_{X^M,Y}}{\widehat{\sigma}^2_{X^M}} \ , \quad \widehat{\beta} = \min\left\{\widehat{s}_0, \max\left\{0, \frac{\widehat{\sigma}_{X^S,Y}}{\widehat{\sigma}^2_{X^S}}\right\}\right\} \text{ and } \widehat{B} = \overline{Y} -_H \left(\widehat{\alpha}\overline{X^M} + \widehat{\beta}\overline{X^S}\right) \ ,$$

where $\widehat{s}_0 = \min\{\text{spr}Y_i/\text{spr}X_i : \text{spr}X_i \neq 0\}$ ($\widehat{s}_0 = \infty$ if $\text{spr}X_i = 0$ for all $i = 1, \ldots, n$).

4. Simultaneous confidence regions for the regression function

Let $X, Y : \Omega \to \mathcal{K}_c(\mathbb{R})$ be random intervals verifying a linear model (2). Thus, from (3) it is obtained that $E(Y|X = x) = \alpha x^M + \beta x^S + B \in \mathcal{K}_c(\mathbb{R})$, for each $x \in \text{Im}(X)$. Given a desirable confidence level $1 - \rho$, the aim is to develop $(1 - \rho)-$simultaneous confidence regions (or, equivalently, a $(1 - \rho)-$confidence band) for $\{E(Y|X = x) : x \in Im(X)\}$. As usual, the regions will be centered around the corresponding estimate $\widehat{E}(Y|X = x)$, and their radii will be computed in terms of the metric d_θ.

Let $\{X_i, Y_i\}_{i=1}^n$ be a random sample obtained from (X, Y), and let $\rho \in (0, 1)$. For each $x \in \text{Im}(X)$, the $(1 - \rho)-$confidence region for $E(Y|X = x)$ is defined as

$$(5) \quad B_{\widehat{E}(Y|X=x)}(\widehat{\delta}_x) = \left\{A \in \mathcal{K}_c(\mathbb{R}) : d_\theta(A, \widehat{E}(Y|X = x)) \leq \widehat{\delta}_x\right\} \ ,$$

where $\{\widehat{\delta}_x : x \in \text{Im}(X)\}$ are computed by assuring the simultaneous coverage condition, that is:

$$P\Big(E(Y|X = x) \in B_{\widehat{E}(Y|X=x)}(\widehat{\delta}_x), \forall x \in \text{Im}(X)\Big) = 1 - \rho \ .$$

Taking inspiration on the classical linear regression problems, the radii in (5) are defined by means of the expression

$$\widehat{\delta}_x = \widehat{K}\sigma_\varepsilon\sqrt{\widehat{\sigma}^2_X + d^2_\theta(x, \overline{X})} \ ,$$

for a suitable constant $\widehat{K}$, for all $x \in \text{Im}(X)$. From the simultaneous coverage condition, it can be obtained:

$$\begin{aligned} 1 - \rho &= P\Big(E(Y|X = x) \in B_{\widehat{E}(Y|X=x)}(\widehat{\delta}_x), \forall x \in \text{Im}(X)\Big) \\ &= P\left(d_\theta\Big(E(Y|X = x), \widehat{E}(Y|X = x)\Big) \leq \widehat{K}\sigma_\varepsilon\sqrt{\widehat{\sigma}^2_X + d^2_\theta(x, \overline{X})}, \forall x\right) \\ &= P\left(\sup_{x\in\text{Im}(X)} \frac{d_\theta\big(E(Y|X = x), \widehat{E}(Y|X = x)\big)}{\sigma_\varepsilon\sqrt{\widehat{\sigma}^2_X + d^2_\theta(x, \overline{X})}} \leq \widehat{K}\right) \end{aligned}$$

Thus, $\widehat{K}$ can be computed as the $(1 - \rho)-$quantile of the distribution of the statistic

$$\widehat{A}_n = \sup_{x\in\text{Im}(X)} \frac{d_\theta\big(E(Y|X = x), \widehat{E}(Y|X = x)\big)}{\sigma_\varepsilon\sqrt{\widehat{\sigma}^2_X + d^2_\theta(x, \overline{X})}} \ .$$

The exact distribution of $\widehat{A}_n$ is unknown, since no parametric models for the random intervals are defined. Moreover, the asymptotic distribution of $\widehat{A}_n$ seems difficult to get. Alternatively, applying a bootstrap procedure, the value of the constant $\widehat{K}$ can be approximated by the corresponding $(1-\rho)-$quantile of a bootstrap distribution of $\widehat{A}_n$.

In practice, the computation of $\widehat{K}$ can be done by means of the following bootstrap algorithm:

Algorithm:

Let $\{X_i, Y_i\}_{i=1}^n$ be a random sample obtained from (X, Y). Let ρ be a fixed significance level and $B \in \mathbb{N}$ large enough.

P1. Compute the parameter estimates $(\widehat{\alpha}, \widehat{\beta}, \widehat{B})$ and $\widehat{\sigma}_\varepsilon$. Thus, $\widehat{E}(Y|X=x) = \widehat{\alpha}x^M + \widehat{\beta}x^S + \widehat{B}, x \in \text{Im}(X)$.

P2. Generate B bootstrap samples $\{X_i^*, Y_i^*\}_{i=1}^n$ of size n, resampling with replacement from the original sample $\{X_i, Y_i\}_{i=1}^n$.

P3. For each iteration $b = 1, \ldots, B$, compute the parameter estimates $(\widehat{\alpha}^*, \widehat{\beta}^*, \widehat{B}^*)$ from the corresponding bootstrap sample. Thus, $\widehat{E}^*(Y|X=x) = \widehat{\alpha}^* x^M + \widehat{\beta}^* x^S + \widehat{B}^*, x \in \text{Im}(X)$. Compute $\overline{X^*}$, $\widehat{\sigma}^2_{X^*}$, and the value of the bootstrap version of $\widehat{A}_n$:

$$\widehat{A}_n^{*(b)} = \sup_{x \in \text{Im}(X)} \frac{d_\theta\big(\widehat{E}(Y|X=x), \widehat{E}^*(Y|X=x)\big)}{\widehat{\sigma}_\varepsilon \sqrt{\widehat{\sigma}^2_{X^*} + d^2_\theta(x, \overline{X^*})}}$$

P4. Approximate $\widehat{K}$ by the $(1-\rho)-$quantile of the empirical distribution $\{\widehat{A}_n^{*(b)}\}_{b=1}^B$.

5. Simulation studies and practical illustration

The empirical behaviour of the bootstrap algorithm can be shown by means of some simulation studies. Let us define a theoretical situation for two random intervals X and Y associated by means of a linear model (2). Two linear models with the structure of (2) will be investigated.

- **Model $\mathbf{M_1}$**: Let mid$X \sim N(0,1)$, spr$X \sim \chi^2_1$, and the interval error term defined by mid$\varepsilon \sim N(0,1)$ and spr$\varepsilon \sim \chi^2_1$, independent from the previous ones. Let Y be a random interval defined by means of the linear model

  (6) $\quad Y = X^M + X^S + \varepsilon$.

- **Model $\mathbf{M_2}$**: Considering the same interval error ε from above, let X be in this case parametrized by mid$X \sim N(0,1)$ and spr$X = (\text{mid}X)^2$. Interval Y is then defined as

  (7) $\quad Y = (-2)X^M + 3X^S + \varepsilon$.

  In this situation, a certain degree of dependence between mid$X$ and spr$X$ is allowed, so intervals X^M and X^S in (7) are dependent.

The empirical coverage of the simultaneous confidence regions for the regression function of M_1 and M_2 is computed by simulating $k = 10,000$ random samples from the corresponding model for different sample sizes n and running the bootstrap algorithm for each sample. In each iteration of the algorithm, a value of $\widehat{K}$ is computed, and the membership of the regression function to the corresponding confidence region is checked. In Table 1 the simulation results for three nominal significance levels ($100(1-\rho)\% = 90, 95, 99\%$) are presented. In all the cases, the coverage rates approximate the nominal significance level as the sample size increases, which shows the empirical correctness of the bootstrap procedure.

Table 1: Empirical confidence level of the bootstrap simultaneous confidence regions

Model	$n \setminus 100(1-\rho)$	90	95	99
M_1	10	91.58	96.49	99.66
	30	90.48	95.80	99.21
	50	90.47	95.53	99.17
	100	90.01	95.08	98.87
	300	90.07	94.96	98.94
M_2	10	93.39	97.58	99.79
	30	91.59	96.55	99.57
	50	90.43	95.53	99.21
	100	90.11	95.41	99.14
	300	90.09	95.07	99.01

The practical application of the algorithm is illustrated over a real-life data set. Data in Table 2 correspond to the fluctuations over a day of the systolic and diastolic blood pressure of 59 patients who were hospitalized on the hospital *Valle del Nalón*, located in Asturias (Spain). The complete data set can be seen in Gil et al., 2007.

Table 2: Systolic (X) and diastolic (Y) blood pressure fluctuation of some patients

X	Y	X	Y	X	Y
[11.8,17.3]	[6.3,10.2]	[11.9,21.2]	[4.7,9.3]	[9.8,16.0]	[4.7,10.8]
[10.4,16.1]	[7.1,10.8]	[12.2,17.8]	[7.3,10.5]	[9.7,15.4]	[6.0,10.7]
[13.1,18.6]	[5.8,11.3]	[12.7,18.9]	[7.4,12.5]	[8.7,15.0]	[4.7,8.6]
...	...	...	...	...	...

The aim is to express Y by means of a linear model with the structure of (2) in terms of the values of X, and to compute a simultaneous confidence region for that relationship at significance level 95%. From the available data set, the estimated model is $\widehat{Y} = 0.4527X^M + 0.2641X^S + [0.1418, 3.2422]$. If we run the algorithm with $B = 1000$ and $1-\rho = 0.95$, $\widehat{K} = 0.1430$ is obtained. Thus, the $0.95-$simultaneous confidence region for $\{E(Y/X = x) : x \in \mathrm{Im}(X)\}$ has the expression:

$$SCR_{0.95} = \Bigg\{ A \in \mathcal{K}_c(\mathbb{R}) :$$

$$d_\theta\big(A, \widehat{E}(Y|X = x)\big) \leq (0.143)(0.949)\sqrt{3.582 + (\mathrm{mid}x - 14.678)^2 + (\mathrm{spr}x - 3.489)^2}, \forall x \in \mathrm{Im}X \Bigg\} .$$

6. Concluding remarks

A bootstrap algorithm to construct the $(1-\rho)-$simultaneous confidence regions for the regression function of a flexible linear model between random intervals has been proposed. The algorithm has been shown to be empirically correct, since the coverage rates obtained for different simulated data approximate the desirable nominal level in all the investigated situations. Thus, the algorithm has been applied over a real data set in order to show its applicability in practice. It remains the possible development of the procedure by means of an asymptotic approach, studying the limit distribution of the statistic $\widehat{A}_n$. It can be remarked that the asymptotic techniques provide accurate results for very large samples sizes, in general, whereas bootstrap methods are generally more adequate for small or moderate sample sizes.

REFERENCES

Aumann, R.J., 1965. Integrals of set-valued functions. Journal of Mathematical Analysis and Applications 12, 1-12.

Bertoluzza, C., Corral, N., Salas A., 1995. On a new class of distances between fuzzy numbers. Mathware & Soft Computing 2, 71–84.

Blanco-Fernández, 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. et al. (Eds.) Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Intelligent and Soft Computing 77, 33-40.

Blanco-Fernández, A., Corral, N., González-Rodríguez, G., 2011. Estimation of a flexible simple linear model for interval data based on the set arithmetic. Computational Statistics & Data Analysis. *In press.*

Colubi, A., 2009. Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data. Fuzzy Sets and Systems 160 (3), 344-356.

Efron, B., Tibshirani, R., 1993. An introduction to the Bootstrap. Chapman and Hall, New York.

Gil, M.A., González-Rodríguez, G., Colubi, A. Montenegro M., 2007. Testing linear independence in linear models with interval-valued data. Computational Statistics & Data Analysis 51, 3002-3015.

Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences 179 (23), 3964-3972.

Acknowledgements

The research in this paper has been partially supported by the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02. Their financial support is gratefully acknowledged.