A Linearity Test for a Simple Regression Model with *LR* Fuzzy Response

Maria Brigida Ferraro, Ana Colubi, and Paolo Giordani

Abstract. A linearity test for a simple regression model with an imprecise response is investigated. The values of the imprecise response are formalized through *LR*-fuzzy numbers, and the stochastic variability through probability spaces. The linear regression model and the least squares estimators of the regression parameters are briefly recalled. The nonparametric model to be employed as reference in the testing approach is also presented. The statistic compares the variability explained by the linear regression with the one explained by the nonparametric regression, since in case of linearity, both quantities should be similar. The problem is approached by bootstrapping. A simulation study has been carried out in order to check the performance of the procedure.

Keywords: Fuzzy random variable, Fuzzy regression, Linearity test, Bootstrap approach.

1 Introduction

To formalize an imprecise value, a useful kind of fuzzy numbers is the so-called *LR* family. A linear regression model with an *LR* fuzzy response and a real explanatory variable has been introduced and analyzed in [5, 6].

Maria Brigida Ferraro and Paolo Giordani
Dipartimento di Statistica, Prob. e Stat. Applicate - SAPIENZA
Università di Roma, 00185 Roma, Italy
e-mail: `mariabrigida.ferraro@uniroma1.it,paolo.giordani@uniroma1.it`

Ana Colubi
Departamento de Estadística e I.O. y D.M., Universidad de Oviedo,
33007 Oviedo, Spain
e-mail: `colubi@indurot.uniovi.es`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 263–271.
springerlink.com

Among the inferential procedures in a linear regression context, it can be interesting to check the adequacy of the linear regression for modelling the relationship between the imprecise response and the explanatory variable. For this purpose, it is possible to use an expert criterion or an hypothesis test. The aim of this work is to suggest a test statistic to check the linearity of the relationship and its empirical behaviour.

The proposed linearity test takes inspiration from Azzalini & Bowman [1], who suggest to check the linearity of the relationship by comparing the residuals of the linear regression with those resulting from a nonparametric model. Here, we apply this idea in the context of the regression model with *LR* response taking into account the model in [5]. The hypothesis testing problem is approached by bootstrapping. In details, we propose a residual bootstrap test to check the linearity of the relationship.

In the next section we introduce some preliminary concepts. In Section 3 a linear regression model with *LR* fuzzy response and the estimation problem are recalled, and a nonparametric model is presented. Section 4 deals with the proposed linearity test and, in order to check its performance, simulation studies and a real-life example are carried out in Section 5. Finally, Section 6 contains some concluding remarks.

2 Preliminaries

A fuzzy set $\widetilde{A}$ is identified by the *membership function* $\mu_{\widetilde{A}} : \mathbb{R} \to [0,1]$ so that $\mu_{\widetilde{A}}(x)$ is the membership degree of x in the fuzzy set $\widetilde{A}$ [9]. A particular class of fuzzy sets very useful in practice is the *LR* family, $\mathscr{F}_{LR}$, whose members are the so-called *LR fuzzy numbers*, determined by three values: the center, the left and the right spread (see, for example, [2, 3]). Namely, a mapping $s : \mathscr{F}_{LR} \to \mathbb{R}^3$, i.e., $s(\widetilde{A}) = s_{\widetilde{A}} = (A^m, A^l, A^r)$ (where A^m, $A^l \geq 0$, $A^r \geq 0$ are, respectively, the center, the left and the right spread), is associated to each *LR* fuzzy set $\widetilde{A}$. In what follows it is indistinctly used $\widetilde{A} \in \mathscr{F}_{LR}$ or $(A^m, A^l, A^r) \in \mathbb{R}^3$. The membership function of $\widetilde{A} \in \mathscr{F}_{LR}$ can be written as

$$\mu_{\widetilde{A}}(x) = \begin{cases} L\left(\frac{A^m - x}{A^l}\right) & x \leq A^m,\ A^l > 0, \\ 1_{\{A^m\}}(x) & x \leq A^m,\ A^l = 0, \\ R\left(\frac{x - A^m}{A^r}\right) & x > A^m,\ A^r > 0, \\ 0 & x > A^m,\ A^r = 0, \end{cases} \tag{1}$$

where the functions L and R are particular decreasing shape functions from $\mathbb{R}^+$ to $[0,1]$ such that $L(0) = R(0) = 1$ and $L(x) = R(x) = 0, \forall x \in \mathbb{R} \setminus [0,1]$, and 1_I is the indicator function of a set I. $\widetilde{A}$ is a *triangular* fuzzy number if $L(z) = R(z) = 1 - z$, for $0 \leq z \leq 1$.

The operations considered in $\mathscr{F}_{LR}$ are the natural extensions of the Minkowski sum and the product by a positive scalar for intervals. In details, the sum of $\widetilde{A}$ and $\widetilde{B}$ in $\mathscr{F}_{LR}$ is the *LR* fuzzy number $\widetilde{A}+\widetilde{B}$ so that

$$(A^m,A^l,A^r)+(B^m,B^l,B^r)=(A^m+B^m,A^l+B^l,A^r+B^r),$$

and the product of $\widetilde{A}\in\mathscr{F}_{LR}$ by a positive scalar γ is

$$\gamma(A^m,A^l,A^r)=(\gamma A^m,\gamma A^l,\gamma A^r).$$

Yang & Ko [8] define a distance between two *LR* fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$ as follows

$$\begin{aligned}D^2_{LR}(\widetilde{A},\widetilde{B})=(A^m-B^m)^2\;&+\;[(A^m-\lambda A^l)-(B^m-\lambda B^l)]^2\\&+\;[(A^m+\rho A^r)-(B^m+\rho B^r)]^2,\end{aligned}$$

where the parameters $\lambda=\int_0^1 L^{-1}(\omega)\mathrm{d}\omega$ and $\rho=\int_0^1 R^{-1}(\omega)\mathrm{d}\omega$ are related to the shape of the membership function. In the triangular case, $\lambda=\rho=\frac{1}{2}$ (see, for more details, [8]). In order to embed the space $\mathscr{F}_{LR}$ into $\mathbb{R}^3$ by preserving the metric a generalization of the Yang and Ko metric has been derived in [5]. Namely, given $a=(a_1,a_2,a_3)$ and $b=(b_1,b_2,b_3)\in\mathbb{R}^3$, it is

$$D^2_{\lambda\rho}(a,b)=(a_1-b_1)^2+((a_1-\lambda a_2)-(b_1-\lambda b_2))^2+((a_1+\rho a_3)-(b_1+\rho b_3))^2,$$

where λ, $\rho\in\mathbb{R}^+$. According to Puri & Ralescu's sense, the concept of fuzzy random variable (FRV) can be introduced. Let $(\Omega,\mathscr{A},P)$ be a probability space, a mapping $\widetilde{X}:\Omega\to\mathscr{F}_{LR}$ is an *LR* FRV if the *s*-representation of $\widetilde{X}$, $(X^m,X^l,X^r):\Omega\to\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^+$ is a random vector [7]. The expectation of an *LR* FRV $\widetilde{X}$ is the unique fuzzy set $E(\widetilde{X})$ ($\in\mathscr{F}_{LR}$) such that $(E(\widetilde{X}))_\alpha=E(X_\alpha)$ provided that $E\|\widetilde{X}\|^2_{LR}=E(X^m)^2+E(X^m-\lambda X^l)^2+E(X^m+\rho X^r)^2<\infty$, where X_α is the α- level of fuzzy set $\widetilde{X}$, that is, $X_\alpha=\left\{x\in\mathbb{R}|\mu_{\widetilde{X}}(x)\geq\alpha\right\}$, for $\alpha\in(0,1]$, and $X_0=cl(\left\{x\in\mathbb{R}|\mu_{\widetilde{X}}\geq 0\right\})$. In this particular case it results $s_{E(\widetilde{X})}=(E(X^m),E(X^l),E(X^r))$. The variance of $\widetilde{X}$ can be defined as

$$\sigma^2_{\widetilde{X}}=var(\widetilde{X})=E[D^2_{LR}(\widetilde{X},E(\widetilde{X}))]$$

and the covariance between two *LR* FRVs $\widetilde{X}$ and $\widetilde{Y}$ is

$$\begin{gathered}\sigma_{\widetilde{X},\widetilde{Y}}=cov(\widetilde{X},\widetilde{Y})=E\langle s_{\widetilde{X}}-s_{E(\widetilde{X})},s_{\widetilde{Y}}-s_{E(\widetilde{Y})}\rangle_{LR}\\=E((X^m-EX^m)(Y^m-EY^m))\\+E((X^m-EX^m-\lambda(X^l-EX^l))(Y^m-EY^m-\lambda(Y^l-EY^l)))\\+E((X^m-EX^m+\rho(X^r-EX^r))(Y^m-EY^m+\rho(Y^r-EY^r))).\end{gathered}$$

3 A Linear Regression Model and a Nonparametric Model with *LR* Fuzzy Random Response and Real Explanatory Variables

Consider a random experiment in which an *LR* fuzzy response variable $\widetilde{Y}$ and a real explanatory variable X are observed on n statistical units, $\{\widetilde{Y}_i, X_i\}_{i=1,\ldots,n}$. Since $\widetilde{Y}$ is characterized by three real-valued random variables (Y^m, Y^l, Y^r), the regression model proposed in [5] concerns this tuple. The center Y^m can be related to the explanatory variable X through a classical regression model. Due to some difficulties entailed by the non-negativity condition of Y^l and Y^r, the authors proposed to model a transform of the left spread and a transform of the right spread of the response through simple linear regressions (on the explanatory variable X). This can be represented in the following way, letting $g : (0, +\infty) \longrightarrow \mathbb{R}$ and $h : (0, +\infty) \longrightarrow \mathbb{R}$ be invertible:

$$\begin{cases} Y^m = a_m X + b_m + \varepsilon_m, \\ g(Y^l) = a_l X + b_l + \varepsilon_l, \\ h(Y^r) = a_r X + b_r + \varepsilon_r, \end{cases} \tag{2}$$

where ε_m, ε_l and ε_r are real-valued random variables with $E(\varepsilon_m|X) = E(\varepsilon_l|X) = E(\varepsilon_r|X) = 0$. Concerning the spreads, model (2) is linear in the transformed scales represented by functions g and h.

The variance of the explanatory variable X is denoted by σ_X^2 and Σ stands for the covariance matrix of $(\varepsilon_m, \varepsilon_l, \varepsilon_r)$, whose variances are strictly positive and finite. In the sequel we will assume the existence of all population variances and covariances involved in the developments.

In general, an *LR* fuzzy random variable $\widetilde{Y}$ and a (real-valued) random variable X can also be related by means of a nonparametric model. As in (2) we consider jointly three equations in which the response variables are the center Y^m and two transforms of the left and the right spreads ($g(Y^l)$ and $h(Y^r)$) of $\widetilde{Y}$, that is,

$$\begin{cases} Y^m = f_m(X) + \varepsilon_m, \\ g(Y^l) = f_l(X) + \varepsilon_l, \\ h(Y^r) = f_r(X) + \varepsilon_r. \end{cases} \tag{3}$$

To estimate model (2), a least squares (LS) approach has been employed. Let $\widetilde{Y}$ and X be two (fuzzy and real-valued) random variables satisfying model (2) observed on n statistical units, $\{\widetilde{Y}_i, X_i\}_{i=1,\ldots,n}$. It can be shown that the LS estimators for the parameters of model (2) are strongly consistent and their expressions in terms of the sample moments are (see [5])

$$\widehat{a}_m = \frac{\widehat{\sigma}_{XY^m}}{\widehat{\sigma}_X^2}, \quad \widehat{a}_l = \frac{\widehat{\sigma}_{Xg(Y^l)}}{\widehat{\sigma}_X^2}, \quad \widehat{a}_r = \frac{\widehat{\sigma}_{Xh(Y^r)}}{\widehat{\sigma}_X^2}, \quad \widehat{b}_m = \frac{\sum\limits_{i=1}^{n} Y_i^m}{n} - \widehat{a}_m \frac{\sum\limits_{i=1}^{n} X_i}{n},$$

$$\widehat{b}_l = \frac{\sum\limits_{i=1}^{n} g(Y_i^l)}{n} - \widehat{a}_l \frac{\sum\limits_{i=1}^{n} X_i}{n}, \quad \widehat{b}_r = \frac{\sum\limits_{i=1}^{n} h(Y_i^r)}{n} - \widehat{a}_r \frac{\sum\limits_{i=1}^{n} X_i}{n}.$$

Concerning model (3), the functions f_m, f_l and f_r can be estimated in practice by means of nonparametric smoothing. Following [1], a kernel approach can be used yielding

$$\begin{cases} \widehat{f}_m(Z) = \dfrac{\sum\limits_{i=1}^{n} Y_i^m K((Z-X_i)/w)}{\sum\limits_{i=1}^{n} K((Z-X_i)/w)}, \\ \widehat{f}_l(Z) = \dfrac{\sum\limits_{i=1}^{n} g(Y_i^l) K((Z-X_i)/w)}{\sum\limits_{i=1}^{n} K((Z-X_i)/w)}, \\ \widehat{f}_r(Z) = \dfrac{\sum\limits_{i=1}^{n} h(Y_i^r) K((Z-X_i)/w)}{\sum\limits_{i=1}^{n} K((Z-X_i)/w)}, \end{cases} \tag{4}$$

where $K\left(\frac{Z-X_i}{w}\right)$ is a kernel function and w the smoothing parameter. In this case we have used the same w for the three regression models because our aim is not to estimate such a parameter. Nonetheless, in general, three different smoothing parameters can also be considered.

For both the models, the residual sum of squares can be defined as

$$SSE = \sum_{i=1}^{n} D_{\lambda\rho}^2(\widetilde{Y}_i^T, \widehat{\widetilde{Y}^T}), \tag{5}$$

where $\widetilde{Y}_i^T = (Y_i^m, g(Y_i^l), h(Y_i^r))$ and $\widehat{\widetilde{Y}_i}^T = (\widehat{Y}_i^m, \widehat{g(Y_i^l)}, \widehat{h(Y_i^r)})$, $i = 1, ..., n$.

4 A Linearity Bootstrap Test

The goal of this section is to test

$$H_0 : \begin{cases} f_m(X) = a_m X + b_m \\ f_l(X) = a_l X + b_l \\ f_r(X) = a_r X + b_r \end{cases} \tag{6}$$

against the alternative

$$H_1 : f_m(X), f_l(X), f_r(X) \text{ are smooth and non-linear functions.}$$

For testing the null hypothesis the following test statistic is used

$$T_n = \frac{SSE_0 - SSE_1}{SSE_1}, \tag{7}$$

where SSE_0 is the residual sum of squares under H_0 according to the model in (2), and SSE_1 is the residual sum of squares according to the model in (3), where $\widehat{\widetilde{Y}^T}_i = (\widehat{Y}_i^m, \widehat{g(Y_i^l)}, \widehat{h(Y_i^r)}) = (\widehat{f}_m(X), \widehat{f}_l(X), \widehat{f}_r(X))$ are the values estimated by means of kernel functions in (4).

Remark 1. We suggest to use a gaussian kernel, that is,

$$K\left(\frac{Z-X_i}{w}\right) = \frac{1}{\sqrt{2\pi}w} exp\left(-\frac{(Z-X_i)^2}{2w^2}\right).$$

In this work we propose to fix the smoothing parameter w. It has been proved that the value of w is expected not to be important since the level of the test is unaffected by this value (see, for instance, [1]). In practice, suitable values of w are from $1/n$ to $1/2$ times the range of the X-values. Nevertheless, the power of the test could be affected by the selection of the smoothing parameter.

A bootstrap approach can be used for testing the linearity. More specifically, we generate B bootstrap samples from a bootstrap population fulfilling the null hypothesis in (6), by means of a residual approach [4]. Then, a standard bootstrap algorithm can be implemented using the bootstrap statistic given by

$$T_n^* = \frac{SSE_0^* - SSE_1^*}{SSE_1^*}.$$

For the sake of convenience, the bootstrap algorithm according to the residual approach is summarized as follows:

Step 1: Compute the values $\widehat{a}_m$, $\widehat{a}_l$, $\widehat{a}_r$, $\widehat{b}_m$, $\widehat{b}_l$, $\widehat{b}_r$ and T_n.

Step 2: Compute the residuals $e_i^m = Y_i^m - \widehat{a}_m X_i - \widehat{b}_m$, $e_i^l = g(Y_i^l) - \widehat{a}_l X_i - \widehat{b}_l$, $e_i^r = h(Y_i^r) - \widehat{a}_r X_i - \widehat{b}_r$.

Step 3: Generate a bootstrap sample of the form

$$\Big\{\Big(X_1, Z_1^m = \widehat{Y}_1^m + e_{i_1}^m, Z_1^l = \widehat{g(X_1^l)} + e_{i_1}^l, Z_1^r = \widehat{h(X_1^r)} + e_{i_1}^r\Big), \ldots,$$
$$\Big(X_n, Z_n^m = \widehat{Y}_n^m + e_{i_n}^m, Z_n^l = \widehat{g(X_n^l)} + e_{i_n}^l, Z_n^r = \widehat{h(X_n^r)} + e_{i_n}^r\Big)\Big\},$$

where $\{i_1, i_2, \ldots, i_n\}$ is a random sample of the integers 1 through n, $\widehat{Y}_i^m = \widehat{a}_m X_i + \widehat{b}_m$, $\widehat{g(X_i^l)} = \widehat{a}_l X_i + \widehat{b}_l$, $\widehat{h(X_i^r)} = \widehat{a}_r X_i + \widehat{b}_r$, $i = 1, \ldots, n$, and compute the value of the bootstrap statistic T_n^*.

Step 4: Repeat Step 3 a large number B of times to get a set of B estimators, denoted by $\{T_{n1}^*, \ldots, T_{nB}^*\}$.

Step 5: Approximate the bootstrap p-value as the proportion of values in $\{T_{n1}^*, \ldots, T_{nB}^*\}$ being greater than T_n.

5 Empirical Results

A simulation experiment has been carried out in order to illustrate the empirical significance of the bootstrap test. Note that we have employed $B = 1000$ replications of the bootstrap estimator and we have considered 10000 iterations of the test at three different nominal significance levels $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$ for different sample sizes n, from 30 to 200. We have randomly generated X behaving as $Unif(-2,2)$, ε_m, ε_l, ε_r as $N(0,1)$ and $Y_m = 3X + 5 + \varepsilon_m$, $Y_2 = g(Y_l) = 1.5X + 3.4 + \varepsilon_l$, $Y_3 = h(Y_r) = 2X + 4.2 + \varepsilon_r$, and we have considered a gaussian kernel with $w = range(X)/n$. The empirical percentages of rejection under H_0 are given in Table 1. It is easy to see that also for small sample sizes n the empirical percentages of rejection are very close to the nominal level. If we consider dependent errors, namely, ε_m behaving as $N(0,1)$ and $\varepsilon_l = \varepsilon_m + \varepsilon_1$, $\varepsilon_r = \varepsilon_m + \varepsilon_2$, with ε_1 and ε_2 behaving as $N(0,0.5)$, we carry out the empirical percentages of rejection under H_0 reported in Table 2. Also in this case, we obtain satisfactory results.

We introduce a real-life example concerning the atmospheric concentration of carbon monoxide (CO) (mg/m^3) and the daily maximum temperature (T) (°C) recorded at "Villa Ada" park in Rome in April, 1-10, 1999 (see Figure 1). The first variable has been managed as a triangular *LR* fuzzy random variable where the center is the mean value of the 24 hourly observations daily recorded, the left spread is given by the deviation of the minimum value from the center and the right spread by the deviation of the maximum value from the center.

In this case we obtain a p-value equal to 0.026, that is, the null hypothesis of linearity should be rejected. Obviously, it should be noted that this result could depend on the choice of the distance, of the kernel function and the smoothing parameter.

Table 1 Empirical percentages of rejection under the hypothesis of linearity.

$n \setminus \alpha \times 100$	1	5	10
30	1.20	5.31	9.94
50	1.17	5.13	10.15
100	1.15	4.82	9.89
200	1.00	4.99	10.20

Table 2 Empirical percentages of rejection under the hypothesis of linearity (dependent errors).

$n \setminus \alpha \times 100$	1	5	10
30	1.14	4.94	9.92
50	1.08	5.12	10.24
100	1.10	5.25	9.94
200	1.12	4.60	9.44

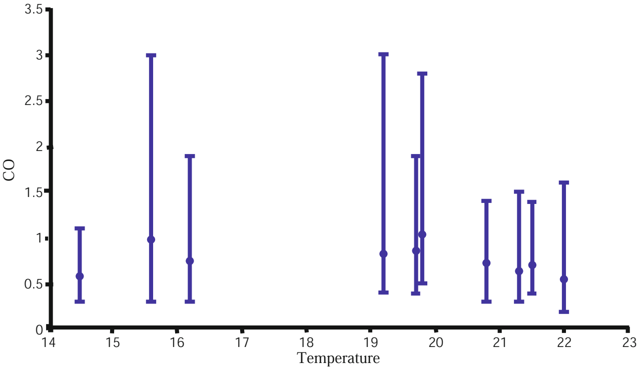

Fig. 1 The observed extreme values of the 0-level and the single-value of CO by the Temperature at "Villa Ada" park in Rome in April, 1-10, 1999

6 Conclusion

In this work we have introduced and analyzed a new linearity test to check the adequacy of a linear relationship between an *LR* fuzzy response and a real explanatory variable. In order to construct a test statistic, we have jointly considered three equations involving the center of the response and two transforms of the left and the right spread and we have taken into account the residual sum of squares based on a suitable distance between *LR* fuzzy numbers. The obtained results are as good as expected in this context. In the near future, it will be interesting to study the power of the test.

References

1. Azzalini, A., Bowman, A.: On the use of nonparametric regression for checking linear relationship. J. R. Statist. Soc. Ser. B 55, 549–557 (1993)
2. Coppi, R., D'Urso, P., Giordani, P., Santoro, A.: Least squares estimation of a linear regression model with LR fuzzy response. Comput. Statist. Data Anal. 51, 267–286 (2006)
3. Di Lascio, L., Ginolfi, L., Albunia, A., Galardi, G., Meschi, F.: A fuzzy-based methodology for the analysis of diabetic neuropathy. Fuzzy Sets Syst. 129, 203–228 (2002)
4. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. Chapman & Hall, New York (1993)
5. Ferraro, M.B., Coppi, R., González-Rodríguez, G., Colubi, A.: A linear regression model for imprecise response. Internat. J. Approx. Reason (2010), doi:10.1016/j.ijar.2010.04.003
6. Ferraro, M.B., Colubi, A., González-Rodríguez, G., Coppi, R.: A determination coefficient for a linear regression model with imprecise response. Environmetrics (accepted for publication, 2010)

7. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
8. Yang, M.S., Ko, C.H.: On a class of fuzzy *c*-numbers clustering procedures for fuzzy data. Fuzzy Sets Syst. 84, 49–60 (1996)
9. Zadeh, L.A.: Fuzzy sets. Inf. Control 8, 338–353 (1965)