

Fuzzy regression and least squares regression: the relationship between two different fitting criteria

*Regressione Fuzzy e Metodo dei Minimi Quadrati:
due differenti metodi di stima*

Rosaria Romano¹ and Francesco Palumbo²

¹ Dipartimento di Matematica e Statistica, Università Federico II di Napoli

² Dipartimento di Istituzioni Economiche e Finanziarie, Università di Macerata
e-mail: francesco.palumbo@unimc.it

Abstract Nell'analisi della dipendenza di una o più variabili quantitative da un set di predittori il metodo dei Minimi Quadrati gioca un ruolo fondamentale. La Teoria delle Probabilità e l'Inferenza classica consentono di valutare l'incertezza delle stime ottenute, mediante la costruzione di opportuni intervalli di confidenza, imponendo al modello ipotesi restrittive spesso difficili da soddisfare. Inoltre esistono dei vincoli non esplicitamente inclusi tra le ipotesi classiche del modello, come quello della multicollinearità tra i predittori, la cui violazione non consente un trattamento "statistico" dell'incertezza delle stime. Il presente lavoro propone la regressione fuzzy come valida alternativa alla regressione statistica in presenza di multi-collinearità. Uno studio empirico basato su simulazioni mostra le differenze tra regressione statistica e fuzzy.

Keywords: fuzzy regression, statistical regression, multi-collinearity.

1. Introduction

Let Y and $\{X\} = \{X_1, X_2, \dots, X_p\}$ be a quantitative dependent variable and a set of independent variables, respectively. Regression analysis studies the dependence of Y with respect to $\{X\}$. There exist many different regression approaches; the most widely used is the (*Multiple*) *Linear Regression*, whose mathematical formalization is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (1)$$

The term ε indicates the deviations between the variable Y and the linear model. The aim is to find the set $\beta = \{\beta_0, \beta_1, \dots, \beta_p\}$ that minimizes the deviations. The β estimator properties and the sample distribution can be specified only if assuming constraints to the model. Under the *Ordinary Least Squares* (OLS) hypotheses, the β estimators have several valuable statistical properties. In some cases these assumptions are too restrictive and their violation affects the estimates because it causes biased and inefficient estimators (Gujarati, 2003).

We focus our attention on the case of “quasi” multi-collinearity. This does not violate any OLS assumption, however it has a dramatic impact on the variance of β , making estimation useless from a practical point of view. In particular, we look for a non classical solution in the so called *fuzzy theory*. Aim of the paper is to present some empirical results obtained from simulation studies showing some remarkable properties of the fuzzy approach in presence of multi-collinearity. The following section shortly

introduces the fuzzy set basic notation and presents the fuzzy regression. In section 3 the main simulation results and a comparison with the OLS approach are illustrated.

2. Fuzzy Regression

Suppose that $Z = \{z_i\}$ is the set of all possible elements with respect to a fuzzy concept. A fuzzy subset A in Z is defined as a set of ordered pairs $\{z_i, \mu_A(z_i)\}$, where $z_i \in Z$, $\mu_A(z_i) : Z \rightarrow [0, 1]$ is the *membership function* of A and $\mu_A(z_i) = h$ is the *grade of membership* of z_i in A . An h -cut (h -level set) of a fuzzy set A in Z is defined as the ordinary set $A_h \subseteq Z$ such that:

$$A_h = \{z_i \in Z : \mu_A(z_i) \geq h\}, \forall h \in [0, 1]$$

The crisp subset of Z consisting of all elements with $\mu_A(z_i) > 0$ is defined *support* of A . If $h = 0$ then $z_i \notin A$, if $h = 1$ then $z_i \in A$ and $0 < h < 1$ when z_i is included in A in the *fuzzy* way. If a fuzzy set $A \subset \mathbb{R}$ satisfies the following properties then it is a *fuzzy number*:

- i) $\sup_{z_i \in \mathbb{R}} \mu_A(z_i) = 1 \longrightarrow$ the fuzzy set is normal;
- ii) $\mu_A[\lambda z_1 + (1 - \lambda)z_2] \geq \mu_A(z_1) \wedge \mu_A(z_2) \longrightarrow$ all h -cuts are convex and bounded, where $\{z_1, z_2\} \in \mathbb{R}$ and $\lambda \in [0, 1]$.

Let us define *interval value* $[x] \equiv [\underline{x}, \bar{x}]$ as an ordered pair of values, where $\underline{x} \leq \bar{x}$. A symmetric triangular fuzzy number is a special case of interval value with linear membership (or possibility) function.

Tanaka's possibilistic fuzzy linear regression model (FR) is defined as (Tanaka et al., 1982):

$$\tilde{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \dots + \tilde{\beta}_j X_j + \dots + \tilde{\beta}_p X_p, \quad (2)$$

where the coefficients are symmetric triangular fuzzy numbers denoted by $\tilde{\beta}_j$. In terms of midpoint and spread $\tilde{\beta}_j = (c_j; a_j)$. Differently from statistical regression, the deviations between the variable Y and the linear model are assumed to depend on the vagueness of the parameters and not on its measurement errors. Tanaka's model minimizes the total spread of the fuzzy coefficients. Spread minimization must be pursued under the constraint of the inclusion of the whole given data set, satisfying a degree of belief h ($0 < h < 1$) chosen by the decision maker. The above analysis leads to a linear programming problem with no restrictive assumptions on the model. Larger values of the h -coefficient imply larger possibility intervals and increase the probability that out-of-sample values will fall within the model. This is comparable to have large confidence intervals in the classical linear regression. Although this model was later improved (Tanaka and Guo, 1999), we focus on it because we are interested to the special case of crisp input/output variables. Our proposal is to compare the classical statistical linear regression model with the FR model in order to evaluate the capabilities of FR in presence of quasi multi-collinearity.

In addition, in order to make a comparison between the Tanaka's approach and the fuzzy least squares approach (FLSR), we propose a comparison between a suitable adaptation of the model proposed by D'Urso and Gastaldi (2000). The model, named *doubly linear adaptive model*, is a model for crisp input and symmetric triangular fuzzy output. It follows the classical approach minimizing the Euclidean distance between the observed and the estimated fuzzy output. Based on the assumption that the spreads are

proportional to the respective centers, first a linear model for the centers is built and after a model for the spreads is derived:

$$\begin{cases} \mathbf{c} = \mathbf{c}^* + \varepsilon_c & \text{where } \mathbf{c}^* = \mathbf{X}\mathbf{a} \\ \mathbf{r} = \mathbf{r}^* + \varepsilon_r & \text{where } \mathbf{r}^* = \mathbf{c}^*b + \mathbf{1}d \end{cases} \quad (3)$$

In the equation (3) $\mathbf{X}$ is the matrix including the independent variables, $\mathbf{c}$ and $\mathbf{c}^*$ are the observed and estimated centers, $\mathbf{r}$ and $\mathbf{r}^*$ are the observed and the estimated spreads, and $\mathbf{a}$, b , d are the parameters of the respective models. Differently from the possibilistic regression, the model (3) does not provide fuzzy parameters. In addition, there is no information about the degree of possibility of the results. However, the estimated output are symmetric triangular fuzzy numbers. In fact in this adaptive procedure the *core model* and the *spread model* produce the centers and spreads of the estimates, respectively. In order to have a model more similar to the possibilistic one, we propose to modify the model (3) in a two-stage least squares model. The model estimation is based on the following two steps:

- i) a classical OLS is performed on the original data;
- ii) a fuzzy dependent variable is defined as $\tilde{Y} = [Y, \hat{e}]$ where Y is the center and $\hat{e}$ is the spread. The spread $\hat{e}$ correspond to the residual from the OLS in the first stage. A FLSM based on the model (3) is then performed.

3. Simulation results and concluding remarks

In order to compare the performance of the ordinary regression and the fuzzy regression models ,in this section a simulation study is carried out. First confidence intervals and fuzzy intervals from Tanaka's model are compared. Then follows a comparison between the Tanaka's approach and the modified model of D'Urso and Gastaldi. Data are simulated under the following hypotheses: two explanatory variables $[X_1, X_2] \sim N(0, 1)$ having mean vector $\mu' = [0, 0]$ and extra diagonal terms of the var-cov matrix $\sigma_{1,2}$ and $\sigma_{2,1}$ equal to 0.6 and 0.7, respectively; 3000 set of 25 statistical units have been used as test data. The confidence and the possibility coefficients have been set equal to 0.95 and 0.75, respectively. For sake of space, we present the results only for one regression coefficient. The model has been simulated both under the null hypothesis of coefficients equal to zero and equal to 0.5. Figure 1[a] and 1[b] show results under hypothesis 1 and 2, respectively. Notice that in both cases intervals of Tanaka's model are narrowest. We have analogous results when the explained variables are simulated collinear. Obviously the OLS intervals are narrower when there is no collinearity among the predictors. These results underline a better performance of the fuzzy regression in presence of high correlations among the predictors. According the two stage procedure proposed in section 2, we also compare the Y estimates, respectively from the OLS, FR, FLSR, "core" models (figure 2[a]). We indicate as "core" model the model built on the centers, with any information on the spreads. Information on the spreads are obtained from figure (2[b]). Looking at the plot (2[a]) it can be soon pointed out that FLSR results are closer to the OLS. Plot (2[b]) shows that FR have spreads wider than FLSR model. This behavior depends on the hypothesis of the linear dependence between the spreads and the centers.

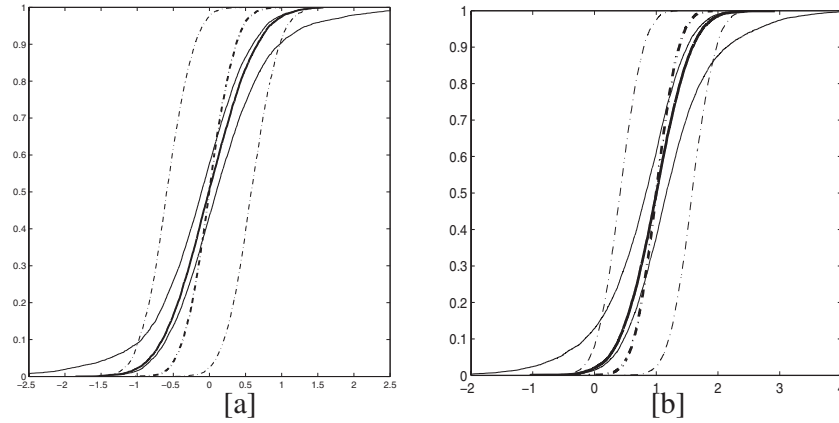


Figure 1: **a)** OLS confidence intervals (dash-dotted lines) and fuzzy intervals (solid lines) under the hypothesis $\beta = 0$; **b)** OLS confidence intervals (dash-dotted lines) and fuzzy intervals (solid lines) under the hypothesis $\beta = 0.5$.

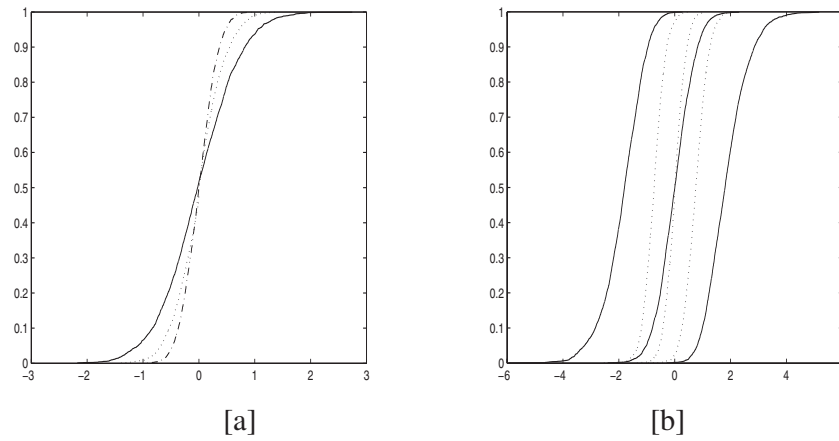


Figure 2: **a)** OLS (dash-dotted lines), FR (solid lines) and FLSR (dotted lines) core models; **b)** FR (solid lines) and FLSR intervals (dotted lines).

References

- D'Urso P., Gastaldi T. (2000) A least-squares approach to fuzzy linear regression analysis, *Computational Statistics and Data Analysis*, 34: 427-440.
- Gujarati D. N. (2003) *Basic Econometrics*, McGraw-Hill, New York.
- Tanaka H., Uejima S., Asai K. (1982) Linear regression analysis with fuzzy model, *IEEE Transactions on SMC*, 12: 903-907.
- Tanaka H., Guo P. (1999) *Possibilistic Data Analysis for Operations Research*, Physica-Verlag, Heidelberg, Germany.
- Zadeh L. A. (1965) Fuzzy sets, *Information and control*, 8(3): 338-353.