The Fuzziness in Regression Models

Arnold F. Shapiro
Smeal College of Business, Penn State University
afs1@psu.edu

Thomas R. Berry-Stölzle
Terry College of Business, University of Georgia
trbs@terry.uga.edu

Marie-Claire Koissi
Department of Mathematics, Western Illinois University
mk122@wiu.edu

ABSTRACT

This article addresses the fuzziness in regression models. The goal is to present a test procedure to explicitly examine whether an independent variable has a clear functional relationship with the dependent variable in a specific regression model, or whether their relationship is fuzzy. To this end, we interpret the spread of the regression coefficients as a statistic measuring the fuzziness of the relationship between the corresponding independent variable and the dependent variable. We then derive test distributions based on the null hypothesis that such spreads could have been obtained with data generated by a classical regression model with random errors. The analysis is presented in conceptual rather than technical terms.

Keywords: fuzzy regression, fuzzy coefficient, test for fuzziness, OLS, simulation

The Fuzziness in Regression Models

1. Introduction

Classical statistical linear regression takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1,2,...,m \quad (1)$$

where the dependent (response) variable, y_i , the independent (explanatory) variables, x_{ij}, and the coefficients (parameters), β_j, are crisp values, and ε_i is a crisp random error term with $E(\varepsilon_i)=0$, variance $\sigma^2(\varepsilon_i)=\sigma^2$, and covariance $\sigma(\varepsilon_i, \varepsilon_j) = 0$, $\forall i,j$, $i \neq j$.

Although statistical regression has many applications, it is problematic if the data set is too small, or there is difficulty verifying that the error is normally distributed, or if there is vagueness in the relationship between the independent and dependent variables, or if there is ambiguity associated with the event, or if the linearity assumption is inappropriate. These are the situations fuzzy regression was meant to address.

In contrast to the classical statistical linear regression, fuzzy regression takes the general form [Tanaka et. al. (1982)]:

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 x_1 + \cdots + \tilde{A}_n x_n \quad (2)$$

where $\tilde{Y}$ is the fuzzy output, $\tilde{A}_i$, i = 0, 1, 2, ..., n, is a fuzzy coefficient, and $\mathbf{x} = (x_1, ..., x_n)$ is an n-dimensional non-fuzzy input vector.

If the fuzzy coefficients are triangular fuzzy numbers (TFNs), their membership functions (MFs), $\mu_A(a)$, can be represented as shown in Figure 1.

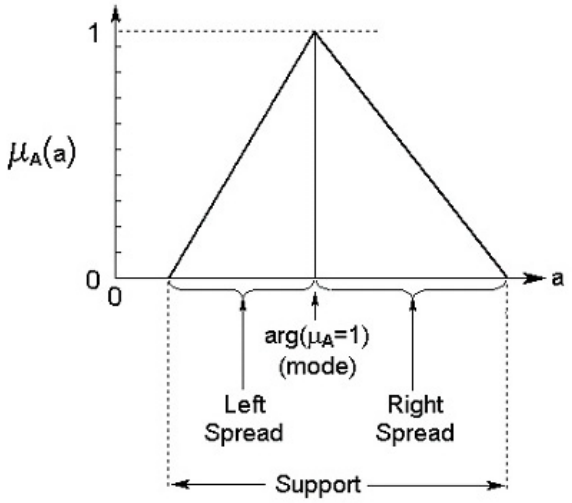

Figure 1: Membership Function for Triangular Fuzzy Number

As indicated, the salient features of the TFN are its mode, its left and right spreads, and its support. When the two spreads are equal, the TFN is known as a symmetrical TFN (STFN).

In this article, which presents a conceptual version of the fuzziness test in Berry-Stölzle et al (2009), we interpret the spread of the regression coefficients as a statistic measuring the fuzziness of the relationship between the dependent variable and the corresponding independent variable. We then derive test distributions based on the null hypothesis that such spreads could have been obtained with data generated by a classical regression model with random errors.

The rest of the article discusses the difference between OLS and fuzzy regression, the fuzziness of the coefficients, the methodology used to test for fuzziness, and the findings of the analysis. We end with comments on the study.

2. Conceptualizing the difference between OLS and fuzzy regression

It is a straightforward matter to conceptualize the essential differences between OLS and fuzzy regression. To this end, we continue an example in Shapiro (2004), portions of which are repeated here.

Consider the following simple Ishibuchi (1992) data:

Table 1: Data Pairs

i	1	2	3	4	5	6	7	8
x_i	2	4	6	8	10	12	14	16
y_i	14	16	14	18	18	22	18	22

Based on this data, OLS results in the regression line shown in Figure 2.

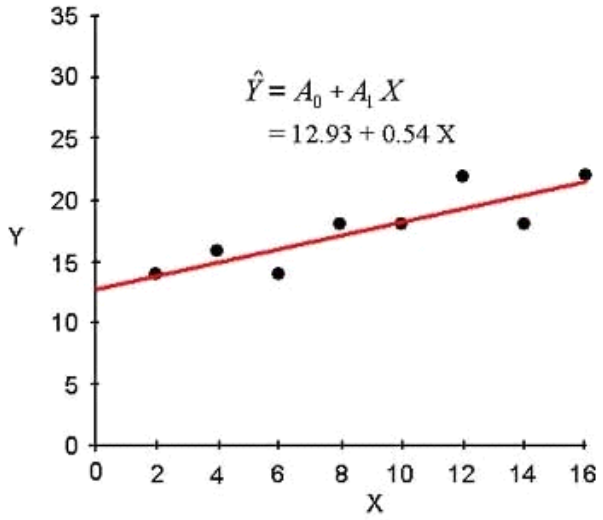

Figure 2: Statistical Linear Regression Example

A comparable fuzzy regression line can be developed using the possibilistic regression (PR) of Tanaka et al (1982). Based on possibilistic distributions, the essential idea of PR is to minimize

the fuzziness of the model by minimizing the total spread of the fuzzy coefficients, subject to including all the given data. While the PR approach has certain drawbacks, it is sufficient for the purpose of this article.[1]

Thus, starting with the Table 1 data, we fit a straight line through two or more data points in such a way that it bounds the data points from above. Here, these points are determined heuristically and OLS is used to compute the parameters of the line labeled Y^U, which takes the values 13 + .75x, as shown in Figure 3(a).

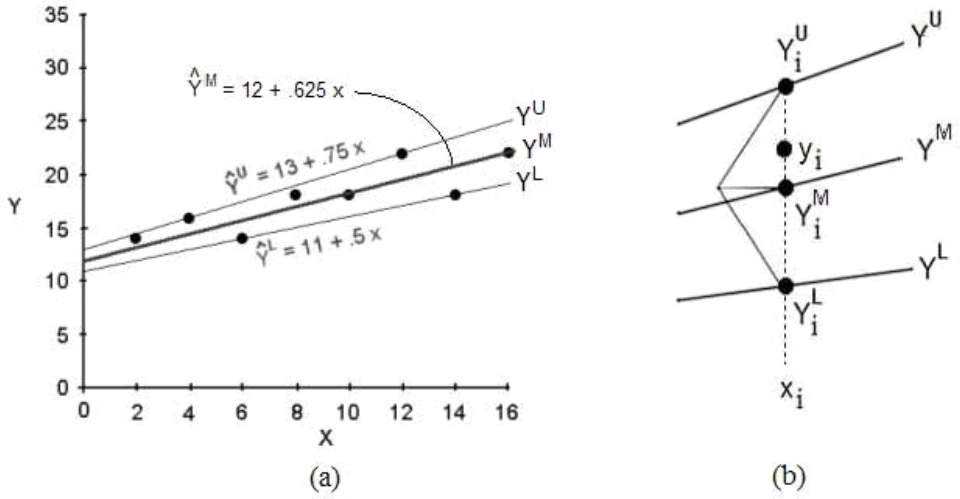

Figure 3: Fuzzy Regression Example

Similarly, we fit a second straight line through two or more data points in such a way that it bounds the data points from below. As indicated, the fitted line in this case, labeled Y^L, takes the values 11 + .5x.

For any given data pair, (x_i, y_i), the foregoing conceptualizations can be summarized by the fuzzy regression interval $[Y_i^L, Y_i^U]$ shown in Figure 3(b).[2] Assuming, for the purpose of this example, that STFN are used for the MFs, the modes of the MFs fall midway between the boundary lines, as shown by the curve labeled Y^M in the figure, that is, $Y_i^M = (Y_i^U + Y_i^L)/2$.

Given the parameters, (Y^U, Y^L, Y^M), which characterize the fuzzy regression model, the i-th data pair (x_i, y_i), is associated with the model parameters.

3. Interpreting the fuzziness of the coefficients

The mere finding that the spread of some of the MF of the coefficients are positive does not necessarily imply a fuzzy relationship between the dependent and independent variables. Unlike OLS, which includes an error term to capture random deviations, possibilistic regression has no

[1] The two main approaches to fuzzy regression are the PR model of Tanaka et al (1982) and the fuzzy least-squares regression models of Diamond (1988). The pros and cons of each are discussed in Diamond and Tanaka (1998) and Shapiro (2004).

[2] Adapted from Wang and Tsaur (2000), Figure 1.

such term. As a consequence, the MFs of some of the coefficients would need to have a positive spread in order to accommodate that variability. Thus, in order to establish a fuzzy relationship between the dependent and independent variables, one has to show that the spread of the MFs exceeds what would be expected simply because of the variability of the OLS error term.

Figure 4 shows a rendition of the situation.

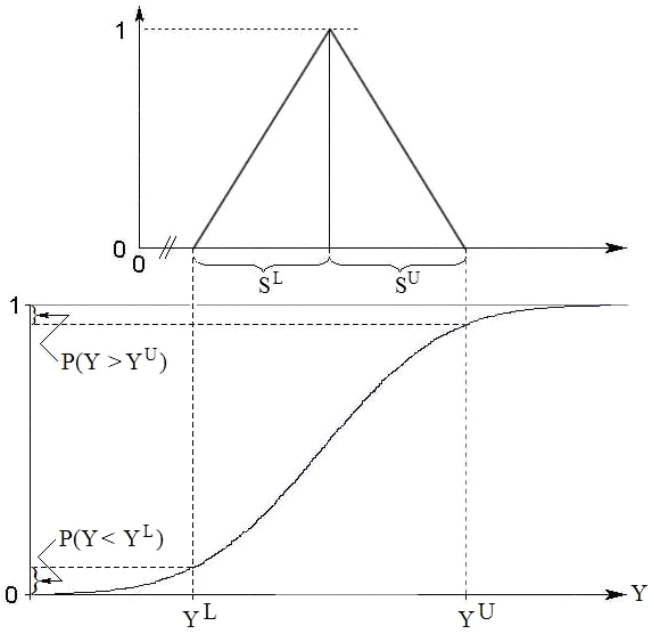

Figure 4: Probability of exceeding the spread

The top curve represents the MF of a fuzzy coefficient, which is taken to be a STFN. As indicated, the right and left spreads of the MF have length S^L and S^U, respectively. The bottom curve shows the probability that the spreads of the MF will be exceeded, given data that conforms to the assumptions of OLS. Thus, as indicated, the probability that S^L and S^U will be exceeded is $P(Y < Y^L)$ and $P(Y > Y^U)$, respectively.

Notice that the proximity of the OLS curve and the mode of the MF of the FR coefficients will determine the relative sizes of the upper and lower probabilities. If the OLS curve passes through roughly the same point as the mode of the MF occupies, the cumulative distribution will seems symmetrical with respect to the MF. In general, however, as represented in the figure, this need not be the case.

4. Methodology

The methodology for our simple example proceeds in three steps. In the first step, we estimate the parameters of the OLS regression model using the Table 1 dataset. This gives us estimates of the coefficients and the empirical standard deviation of the OLS residuals. In the second step, we simulate random OLS output values for each of the independent variable data points. These take the form of $\hat{Y} + \varepsilon^*$, where $\hat{Y}$ denotes the empirical regression line and ε^* is the simulated

error term. In the third step, we compute the empirical probability that the simulated OLS values fall outside the support of the original MFs of the coefficients, that is, 1- $P(Y^L < \hat{Y} + \varepsilon^* < Y^U)$.

5. Findings

Based on a simulation of 80,000 trials (10,000 for each x_i in Table 1), the probability that a random value based on the OLS curve developed from our dataset would fall outside the interval $[Y^L, Y^U]$ is 27 percent. Thus, there is a relatively high probability of obtaining the empirical spreads, and we cannot reject the hypothesis that the relationship between the dependent and independent variable is not fuzzy.

6. Comments

This article discussed a test procedure to explicitly examine whether an independent variable has a clear functional relationship with the dependent variable in a specific regression model, or whether its relationship is fuzzy.

The dataset considered was simplistic and the emphasis was on conceptual rather than technical issues. Consequently, a number of relevant topics were not addressed, such as non normality and heteroscedasticity. Nor did we address the issue of whether the possibilistic regression model provided a better fit to the dataset than the classical regression model with a random error term. These issues and a more general methodology for computing the critical probabilities are addressed in Berry-Stölzle et al (2009).

References

Berry-Stölzle, T. R., M.-C. Koissi and A. F. Shapiro. (2009) "Detecting fuzzy relationships in regression models: the case of insurer solvency surveillance in Germany," Working paper

Diamond, P. (1988) "Fuzzy least squares," Information Sciences 46(3), 141-157

Diamond, P. and H. Tanaka. (1998) “Fuzzy regression analysis,” in R. Stowiński (Ed), Fuzzy sets in decision analysis, operations research, and statistics, Kluwer Academic Publishers, Norwell, Massachusetts

Ishibuchi, H. (1992) "Fuzzy regression analysis," Fuzzy Theory and Systems, 4, 137-148

Shapiro, A. F. (2004) “Fuzzy regression and the term structure of interest rate revisited,” Proceedings of the 14th International AFIR Colloquium, Vol. 1, 29-45.

Tanaka, H., S. Uejima, and K. Asai. (1982) "Linear regression analysis with fuzzy model," IEEE Transactions on Systems, Man and Cybernetics, 12(6), 903-907.

Wang, H.-F. and R.-C. Tsaur. (2000) "Insight of a fuzzy regression model," *Fuzzy Sets and Systems,* 112(3), 355-369