Using Simulation Technique to overcome the multi-collinearity problem for estimating fuzzy linear regression parameters.

Hazim Mansoor Gorgees¹ and Mariam Mohammed Hilal²

¹ Department of Mathematics, College of Education for Pure Science, Ibn-Al-Haitham, University of Baghdad, Iraq hazim5656@yahoo.com
² Department of Mathematics, College of Education for Pure Science, Ibn-Al-Haitham, University of Baghdad, Iraq marmoon80@yahoo.com

Abstract: Fatigue cracking is one of the common types of pavement distresses and is an indicator of structural failure; cracks allow moisture infiltration, roughness, may further deteriorate to a pothole. Some causes of pavement deterioration are: traffic loading; environment influences; drainage deficiencies; materials quality problems; construction deficiencies and external contributors. Many researchers have made models that contain many variables like asphalt content, asphalt viscosity, fatigue life, stiffness of asphalt mixture, temperature and other parameters that affect the fatigue life. For this situation, a fuzzy linear regression model was employed and analyzed by using the traditional methods and our proposed method in order to overcome the multi-collinearity problem. The total spread error was used as a criterion to compare the performance of the studied methods. Simulation program was used to obtain the required results.

Keywords: triangular Fuzzy numbers; Fuzzy linear regression; Least Square method; principle component method and linear programming problem.

1. Introduction:
Regression analysis is a powerful and comprehensive method for analyzing relationships between a response variable (depended variable) and one or more explanatory variables (independent variables). Inferential problems associated with regression model involve the estimation of the model parameters and prediction of response variable from knowledge of explanatory variables, based on a set of crisp data. Moreover, it is usually assumed that the parameters of underlying model are exact numbers (i.e. the relationship between variables is crisp). But in the complex systems, such as the systems existing in biology, agriculture, engineering and economy, we frequently cannot get the exact numerical data for the information of systems because of the complexity of systems themselves, the vagueness in people’s thinking and judgment and the influence of various uncertain factors existing in boundary fuzzy environment around the systems. For this situation, the traditional least squares regression may not be applicable. We need therefore to investigate some soft methods for dealing with these situations. Fuzzy set theory provides suitable tools for regression analysis when the relationship between variables is vaguely defined and the observations are reported as imprecise quantities.

After introducing fuzzy set theory, several approaches to fuzzy regression have been developed by many researchers.

Studies related to fuzzy linear regression may be roughly divided into two approaches, namely, linear programming based methods (possibilistic approach) and fuzzy least squares methods (least squares approach) [2].

The main contribution of this work is to investigate the fuzzy regression model with the existence of multi-collinearity, and suggesting a method to deal with this problem.
2. Fuzzy regression methods:
Several studies related with the fuzzy linear regression started after introducing fuzzy set theory by Zadeh (1961) [6].

At the first time Tanaka et al. (1982) proposed a linear regression model with fuzzy parameters and crisp observed data. Their method has been developed in different directions by some researchers.

Tanaka’s approach is essentially based on transforming the problem of fitting a fuzzy model on a data set to a linear programming problem.

Tanaka et al. regarded a fuzzy data as a possibility distribution. They supposed that the deviations between the observed values and the estimated values are due to the fuzziness of the system structure being investigated. This structure was represented as a fuzzy linear function whose parameters were given by fuzzy sets with membership functions regarded as possibility functions, instead of as probability functions [4].

Fuzzy linear regression models may be classified into three main categories according to the fuzziness of input and output data. Specifically [5]:

i. Each of input and output data are fuzzy.
ii. Input data are crisp but output data are fuzzy.
iii. Each of input and output data are crisp.

The basic model assumes a fuzzy linear function as:

\[ \tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + \cdots + \tilde{\beta}_p x_{ip} \quad , \quad \text{for } i = 1, 2, \ldots, n \]  
(1)

Which can be written as:

\[ (y_i, e_i) = (a_0, c_0) + (a_1, c_1)x_{i1} + (a_2, c_2)x_{i2} + \cdots + (a_p, c_p)x_{ip} \]  
(2)

Where \( \tilde{y}_i = (y_i, e_i) \) the fuzzy output with symmetric triangular numbers, \( y_i \) are the central values and \( e_i \) the spread values.

\( \tilde{\beta}_j = (\alpha_j, c_j) \) are fuzzy coefficients presented in the form of symmetric triangular fuzzy numbers where \( \alpha_j \) denote the central value of the parameter and \( c_j \) are its spread (assuming to be equal from the left and the right), \( c_j \geq 0, j = 1, 2, \ldots, p \).

\( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \) are the vectors of crisp explanatory variables.

In order to analyze the fuzzy linear regression, different approaches were followed. In the following sections, some of these approaches are considered.

3. Tanaka’s model:
In (FLR) analysis, some assumptions concerned traditional regression analyses are relaxed and the uncertainty is represented by a fuzzy relationship between the input and output.

The present paper considers first the model of Tanaka which is a pioneer for such models.

The linear programming formula of the fuzzy regression problem can be written as follows [1]:

\[ \text{Minimize } S = \sum_{i=1}^{n} \sum_{j=0}^{p} c_j |x_{ij}| \]  
(3)

Subject to:

\[ \sum_{j=0}^{p} \alpha_j x_{ij} + (1-h) \sum_{j=0}^{p} c_j |x_{ij}| \geq y_i + (1-h)e_i \]  
(4)
\[ \sum_{j=0}^{p} \alpha_{j}x_{ij} - (1 - h) \sum_{j=0}^{p} c_{j} |x_{ij}| \leq y_{i} - (1 - h)e_{i} \]  \hspace{1cm} (5)

\[ c_{j} \geq 0, \alpha_{j} \in R, \ x_{i0} = 1, \ i = 1, 2, ..., n, \ 0 \leq h \leq 1 \]  \hspace{1cm} (6)

The h value (degree of confidence) is selected by the decision maker, where h is belongs to [0, 1].

4. Savic & Pedrycz approach
Savic and Pedrycz formulated the fuzzy regression by combining the ordinary least squares with minimum fuzziness criterion. The method is constructed in two successive steps. The first step employs ordinary least square regression to obtain fuzzy regression parameters. The minimum fuzziness criterion is used in the second step to find the spread of fuzzy regression parameters.

In the first step, the available information about the value of the center of the fuzzy observations is used to fit a regression line to the data.

In fact, the fuzzy data are regressed as simplified crisp data and the regression analysis is conducted as it is an ordinary least squares regression. The results of this step are employed as center values of the fuzzy regression parameters.

In the next step, the minimum fuzziness criterion is used to determine fuzzy parameters. Spreads of the fuzzy parameters are obtained by equation (4), (5) as the minimum fuzziness method with the distinction of employing the fuzzy centers of regression parameters resulting from the first step [3].

5. Proposed method:
Our proposed method is to deal with case of multi-collinearity among the crisp explanatory variables and it is a modification of Savic-Pedrycz method.

This method summarized as follows:

The principal component is used instead of ordinary least squares regression to determine fuzzy center values of fuzzy regression coefficients in the first step.

The minimum fuzziness criterion is used in the second step to find the spread of fuzzy regression coefficients.

6. Experimental study:
By using MATLAB program we generate a response variable and a set of four explanatory variables each with 54 observations including the two cases represented by absence and existence of the multicollinearity problem among the explanatory variables.

In table 1 the upper and lower bound for each variable in the model is represented.

| Table 1. Upper and lower bound for each variables. |
|-----------------------------------------------|
| Y    | [0.026670837, 0.888949719] |
| X1   | [-8.517193191, -7.35597863] |
| X2   | [8.160232492, 8.803874764] |
| X3   | [-0.798507696, -0.198450939] |
| X4   | [2.114, 7.87] |

Where:

\( x_{1j} \): Initial tensile strain at 5th repetition of bending beam, (the first independent variable).

\( x_{2j} \): Initial flexural stiffness modulus, (the second independent variable).
\( x_3 \): stress level, (the third independent variable).

\( x_4 \): Percent air void, (the fourth independent variable).

**Case 1:** With absence of multi-collinearity:

By using MATLAB programs a simulation model was constructed to determine which of ordinary least square method and principal component method performed well by using MSE as a measure of performance on the data generated with the absence of multi-collinearity problem.

The MSE for the estimated regression coefficients vector is calculated for the two methods of estimation, namely: ordinary least square and principal component methods, the results are presented in table 2.

| Method               | Least Square | Principal Component |
|----------------------|--------------|---------------------|
| MSE                  | 0.072587730448981 | 0.073574164021569 |

It is clear that the MSE obtained by using ordinary least square method is less than the MSE obtained by the principal component method and hence we use it in the first step in the Savic & Pedrycz method. The results are given in table 3.

| Crisp Data | fuzzy data before using Lp | Outputs fuzzy data for Savic & Pedrycz method |
|------------|-----------------------------|-----------------------------------------------|
| \( y \)    | \( \hat{y} \) from Least square | \( y_1=\)centers \( e_1=\) spread | \( \hat{y}_1=\)center \( \hat{e}_1=\)spread |
| 1          | 0.320934739 | 0.530105004 | 0.425519872 | 0.003191399 | 0.425519872 | -0.017816269 |
| 2          | 0.320934739 | 0.420191835 | 0.370563287 | 0.002779225 | 0.370563287 | -0.014434844 |
| 3          | 0.320934739 | 0.490701692 | 0.405819216 | 0.003043637 | 0.405819216 | -0.01787358 |
| 4          | 0.280926490 | 0.427526520 | 0.354226505 | 0.002656699 | 0.354226505 | -0.014349666 |
| 5          | 0.260305654 | 0.632918337 | 0.44661996 | 0.003349590 | 0.44661996 | -0.017298886 |
| 6          | 0.239250631 | 0.428695555 | 0.333970093 | 0.002504776 | 0.333970093 | -0.01525978 |
| 7          | 0.217742741 | 0.442487666 | 0.330115204 | 0.002475864 | 0.330115204 | -0.015420197 |
| 8          | 0.195762075 | 0.383626278 | 0.289694177 | 0.002172706 | 0.289694177 | -0.015275427 |
| 9          | 0.150295937 | 0.413357975 | 0.320815866 | 0.002406119 | 0.320815866 | -0.01428609 |
| 10         | 0.150295937 | 0.39177792 | 0.27103865 | 0.002032769 | 0.27103865 | -0.015398435 |
| 11         | 0.126763423 | 0.415877221 | 0.271320322 | 0.002034902 | 0.271320322 | -0.01392175 |
| 12         | 0.807977244 | 0.663047842 | 0.588781362 | 0.004415860 | 0.588781362 | -0.014050905 |
| 13         | 0.795855349 | 0.630407842 | 0.729451595 | 0.005470887 | 0.729451595 | -0.018441464 |
| 14         | 0.795855349 | 0.65768408 | 0.726768789 | 0.00540827 | 0.726768789 | -0.018500297 |
| 15         | 0.783584708 | 0.566444445 | 0.675014577 | 0.005062609 | 0.675014577 | -0.017380940 |
| 16         | 0.771161626 | 0.597104955 | 0.68413329 | 0.005131000 | 0.68413329 | -0.017595542 |
| 17         | 0.758582267 | 0.553902497 | 0.656242382 | 0.004921818 | 0.656242382 | -0.017662343 |
| 18         | 0.888949719 | 0.446272860 | 0.66761290 | 0.00507085 | 0.66761290 | -0.013837324 |
| 19         | 0.877775964 | 0.609846373 | 0.74381169 | 0.005578584 | 0.74381169 | -0.016330189 |
| 20         | 0.866475944 | 0.610947816 | 0.73871188 | 0.005540339 | 0.73871188 | -0.014749717 |
The regression equation is:

\[ \hat{y}_i = (\ldots) + (0.18802968, 0.0012692, 0.00047394)x_{ij} + (0.037044369, 0.00025005)x_{ij} \]

We found that the total spread is equal to 3.219421578.

Case 2: With existence of multi-collinearity:

As in case 1, we construct a simulation model using MATLAB program in order to determine which of the estimation methods mentioned earlier is the best with the existence of multi-collinearity problem among the explanatory variables. The MSE is employed as an indicator of the performance for each method.
It is obvious that the value of MSE obtained by using principal component method is less than the value of MSE obtained by the least square method. Hence we use it in the first step of Savic & Pedrycz method instead of ordinary least square used in case 1 and this is our proposed method.

The results are shown in table 5.

**Table 5. Results of the proposed method for h=0.5.**

| Crisp Data | fuzzy data before using Lp | Outputs fuzzy data for Savic & Pedrycz method |
|------------|-----------------------------|-----------------------------------------------|
| Y          | ŷ from Least square         | 𝑦̂𝑖=values & 𝑒𝑖=spread                        | 𝑦̂𝑖=center & 𝑒𝑖=spread                        |
| 1          | 0.795855349                 | 0.533955501                                 | 0.664905425                                 | -0.000313579                                 |
| 2          | 0.831788925                 | 0.539194706                                 | 0.685491816                                 | -0.000320426                                 |
| 3          | 0.706620070                 | 0.696683911                                 | 0.701654231                                 | -0.000535689                                 |
| 4          | 0.866475944                 | 0.539482229                                 | 0.700730087                                 | -0.000312889                                 |
| 5          | 0.843485463                 | 0.539053396                                 | 0.691269429                                 | -0.000318504                                 |
| 6          | 0.102663757                 | 0.492721863                                 | 0.297692810                                 | -0.000298454                                 |
| 7          | 0.280926490                 | 0.482235488                                 | 0.381580899                                 | -0.000286444                                 |
| 8          | 0.432167891                 | 0.481470891                                 | 0.456819392                                 | -0.000286136                                 |
| 9          | 0.563511756                 | 0.358909064                                 | 0.461251360                                 | -0.000091028                                 |
| 10         | 0.706620070                 | 0.481310401                                 | 0.593965236                                 | -0.000285413                                 |
| 11         | 0.359403715                 | 0.363570315                                 | 0.361487015                                 | -0.00035136                                 |
| 12         | 0.329343739                 | 0.354302892                                 | 0.337618816                                 | -0.00085384                                 |
| 13         | 0.260305654                 | 0.704372427                                 | 0.481889041                                 | -0.000544755                                 |
| 14         | 0.795853549                 | 0.704910272                                 | 0.750423188                                 | -0.000546512                                 |
| 15         | 0.217742741                 | 0.626542831                                 | 0.422142786                                 | -0.000501723                                 |
| 16         | 0.499994677                 | 0.624612005                                 | 0.562303341                                 | -0.000497840                                 |
| 17         | 0.637623730                 | 0.692686196                                 | 0.665154963                                 | -0.00030677                                 |
| 18         | 0.758582267                 | 0.695628501                                 | 0.727105384                                 | -0.00038158                                 |
| 19         | 0.888949719                 | 0.695770302                                 | 0.792360010                                 | -0.00034973                                 |
| 20         | 0.483464380                 | 0.542207166                                 | 0.512832548                                 | -0.000322047                                 |
| 21         | 0.623234239                 | 0.70451694                                  | 0.663867967                                 | -0.00046002                                 |
| 22         | 0.745842651                 | 0.632608619                                 | 0.689225635                                 | -0.00031029                                 |
| 23         | 0.877775964                 | 0.703805551                                 | 0.790790757                                 | -0.000547298                                 |
| 24         | 0.679590067                 | 0.541732542                                 | 0.610661304                                 | -0.00032112                                 |
| 25         | 0.665790502                 | 0.490908859                                 | 0.578352455                                 | -0.000297864                                 |
| 26         | 0.516256160                 | 0.632774123                                 | 0.574515142                                 | -0.000540822                                 |
| 27         | 0.608346663                 | 0.632498688                                 | 0.620562175                                 | -0.000511252                                 |
| 28         | 0.819953958                 | 0.704461464                                 | 0.762210052                                 | -0.000546353                                 |
| 29         | 0.771161626                 | 0.632619566                                 | 0.701890596                                 | -0.00050946                                 |
| 30         | 0.57870077                  | 0.703900330                                 | 0.641340204                                 | -0.00054693                                 |
The results were agree with the engineering theory beyond the nature of asphalt since each of initial tensile strain, stress level and percent air void have a negative influence on fatigue life, while initial flexural stiffness modulus has positive effect on fatigue life.

Hence, the least squares estimators were used as centered values in the first step of Pedrycz method. Again the results of Savic & Pedrycz method include that the initial tensile strain, stress level and percent air void have a negative influence on fatigue life, while initial flexural stiffness modulus has positive influence on fatigue life.

The regression equation is:

$$\hat{y}_i = (-1.03, 0.0006585036534) + (-0.1752, 0.000954758854)x_{ij} + (0.001, 0.000099591951)x_{3j} + (-0.3, 0.000608505554)x_{3j} + (-0.00231, 0.0000031229914)x_{4j}$$

The value of total spread for this method is 0.09833923.

7. Conclusions

From our experimental study the following Conclusions are pointed out

1. With absence of multi-collinearity problem, it was found that the least squares method is better than principal component method in the sense of MSE as it is shown in table 2. Consequently, the least squares estimators were used as centered values in the first step of Savic & Pedrycz method, whose results were agree with the engineering theory beyond the nature of asphalt, the results yield from applying Savic & Pedrycz method include that the initial tensile strain, stress level and percent air void have a negative influence on fatigue life while initial flexural stiffness modulus has positive influence on fatigue life.

2. With existence of multi-collinearity problem, it was found that the principal component method is better than least squares method in the sense of MSE as it is shown in table 4. Hence, the principal component estimators were used as center values in the first step of Savic & Pedrycz method. Again the results of Savic & Pedrycz method were coincide with the engineering theory beyond the nature of asphalt since each of initial tensile strain, stress level and percent air void have a negative effect on fatigue life, while initial flexural stiffness modulus has positive effect on fatigue life.
Reference:
[1] Akdemir, Hande and Tiryaki, F., 2013, “Using fuzzy linear regression to estimate relationship between forest fires and meteorological conditions”, *Applications and Applied Mathematics, ISSN: 1932-8366*, Vol. 8, Issue 2 PP: 673-683.

[2] Alsoltany, S. N. and Alnaqash, I. A., (2015), “Estimating Fuzzy Linear Regression Model for Air Pollution Predictions in Baghdad City”, *Journal of Al-Nahrain University*, Vol. 18 (2), PP. 157-166.

[3] Savic, A., Dragan and Pedrycz, Witold, 1991, “Evaluation of fuzzy linear regression models”, *Fuzzy Sets and Systems*, Vol. 39, PP. 51-63.

[4] Tanaka, H., Uejima, S. and Asai, K., 1982, "Linear Regression Analysis with Fuzzy Model", *IEEE Transactions on Systems, Man, and Cybernetics*, VOL. SMC-12, NO. 6.

[5] Yang, M. S. and Lin, T. S., (2002), “Fuzzy least-squares linear regression analysis for fuzzy input-output data”, *Fuzzy Sets and Systems* 126 (2002), PP. 389-399.

[6] Zadeh, L. A., 1965, "Fuzzy Sets ", *Academic Press Inc*, 8(3), PP. 338-353.