ON FUZZY REGRESSION ADAPTING PARTIAL LEAST SQUARES

Alper BASARAN[1]
Assistant Professor,
Department of Mathematics
Nigde University Department of Mathematics, Turkey

E-mail: muratalper@yahoo.com

Biagio SIMONETTI [2]
PhD, Professor
Benevento, Department of Analysis of Economic and Social,
University of Sannio, Benevento, Italy

E-mail: simonetti@unisannio.it

Luigi D'AMBRA [3]
PhD, University Professor
Department of Biology
University Federico II" of Naples, Napoli, Italy

E-mail: dambra@unina.it

Abstract:
Partial Least Squared (PLS) regression is a model linking a dependent variable y to a set of X (numerical or categorical) explanatory variables. It can be obtained as a series of simple and multiple regressions of simple and multiple regressions. PLS is an alternative to classical regression model when there are many variables or the variables are correlated. On the other hand, an alternative method to regression in order to model data has been studied is called Fuzzy Linear Regression (FLR). FLR is one of the modelling techniques based on fuzzy set theory. It is applied to many diversified areas such as engineering, biology, finance and so on. Development of FLR follows mainly two paths. One of which depends on improving the parameter estimation methods. This enables to compute more reliable and more accurate parameter estimation in fuzzy setting. Second of which is related to applying these methods to data, which usually do not follow strict assumptions. The application point of view of FLR has not been examined widely except outlier case. For example, it has not been widely examined how FLR behaves under the multivariate case. To overcome such a problem in classic setting, one of the methods that are practically useful is PLS. In this paper, FLR is examined based on application point of view when it has several explanatory variables by adapting PLS.

Keywords : *Fuzzy regression, partial least squares, fuzzy number*

1. Introduction

Fuzzy set theory (FST) was introduced by Zadeh (1978) in order to model uncertainty in linguistic imprecision. Then, this theory draws attention in many diverse fields. One of the easily applicable areas is the subject of modeling such as regression. FLR was first proposed by Tanaka (1982). In the last three decades, FLR was studied by many researchers in terms of improving parameter estimation. Although several researches have been conducted in order to improve more reliable parameter estimations, the issues emerging from modeling several explanatory variables with respect to application have not been widely examined in FLR. Various methods such as PLS by Garthwaite (1994), Principal component analysis are developed to overcome this issue in classic regression. In this paper, PLS is adapted to fuzzy case when the dependent variable and independent variables are crisp.

To illustrate why FLR as an alternative modeling tool is employed, instead of using classic regression, two data sets are employed. In the first data provided by Tanaka and Guo (1999), which is called Houses Data, it is shown that the classic regression failed since some variable that will be explained in Section 4 is inconsistent with the intuition. Tanaka and Guo (1999) suggested that FLR can be used as an alternative technique to model price against five explanatory variables. Then they used linear programming formulation, that will be given in detail in Section 4, to estimate the fuzzy parameters of the independent variables in the FLR model. However, it also fails since the value of some of the parameters are zero. Therefore, the number of independent variables decreases in FLR when the motivation of explaining the price with those variables is aimed . Hence, despite of the fact that FLR suggested by Tanaka and Guo (1999) as an alternative modeling technique, it still has some issues that should be resolved. For this purpose, a very useful technique called PLS is employed to construct new variables that will used in FLR. PLS end up with one variable which is a linear combination of five independent variables. Then, this new constructed variable is used against the price to estimate FLR model. Following the similar steps in the second data set, which is called *Chocolate Data*, we shown that the same problem has existed. Therefore, the technique called PLS used in classic regression can be adapted to FLR when the dependent and independent variables are crisp.

The rest of the paper is organized as follows. Sections 2 and 3 give a brief description about Partial Least Squares Regression and fuzzy regression respectively. Section 4 gives a concrete example why classic regression fails and explains why fuzzy regression as an alternative technique can be used when assumptions are violated and the functional relationship is unknown. Section 5 gives details of the application of combining PLS and FLR. The last section is the conclusions.

2. Partial Least Squares or Projection to Latent Structure

Partial Least Squares Regression (PLS-Regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the **X** (explicative variables) and **y** (response variable) data are projected to new spaces, the PLS family of methods are known as bilinear factor models.

PLS-regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among **X** values. By contrast, standard regression will fail in these cases. The goal of PLS regression is to predict **y** from **X** and to describe their common structure. When **y** is a vector and **X** is full rank, this goal could be accomplished using multiple regression. When the number of predictors is large compared to the number of observations, **X** is likely to be singular and the regression approach is no longer feasible (i.e., because of multicollinearity). Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors (e.g., using stepwise or forward methods) another one, called Principal Component Regression, is to perform a Principal Component Analysis (PCA) of the **X** matrix and then use the principal components of **X** as regressors on **y**. The orthogonality of the principal components overcomes the multicollinearity problem. But, the problem of choosing an optimum subset of components remains. Different approaches had been proposed in the past to select the optimal number of PCs (Valle et al, 1999): Akaike information criterion, minimum description length, imbedded error function, cumulative percent variance, scree test on residual percent variance, average eigenvalue, parallel analysis, autocorrelation, cross validation based on the PRESS and R-ratio and variance of the reconstruction error.

Following one of the cited methods, it is possible to keep only a few of the first components.

But they are chosen to explain **X** rather than **y**, and so, nothing guarantees that the principal components, which "explain" **X**, are relevant for **y**.

By contrast, PLS regression searches for a set of components (called latent vectors) that performs a simultaneous decomposition of **X** and **y** with the constraint that these components explain as much as possible of the covariance between **X** and y. This step generalizes PCA. It is followed by a regression step where the decomposition of **X** is used to predict **y**. Simultaneous decomposition of predictors and dependent variables PLS regression decomposes both **X** and **y** as a product of a common set of orthogonal factors and a set of specific loadings. So, the independent variables are decomposed as **TP**′ where **T** and **P** are the score and loadings matrices respectively with **T′T**=**I** with **I** being the identity matrix. By analogy with PCA, **T** is called the score matrix, and **P** the loading matrix (in PLS regression the loadings are not orthogonal). The columns of **T** are the latent vectors. When their number is equal to the rank of **X**, they perform an exact decomposition of **X**.

2.1 PLS regression and covariance

The latent vectors could be chosen in a lot of different ways. In fact in the previous formulation, any set of orthogonal vectors spanning the column space of **X** could be used to play the role of **T**. In order to specify **T**, additional conditions are required. For PLS regression this amounts to finding two sets of weights *w* and *c* in order to create (respectively) a linear combination of the columns of **X** and **y** such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors **t**=**Xw** and **u**=**Yc** with the constraints that **w**′**w** =1, **t**′**t**=1 and **t**′**u** be maximal.

When the first latent vector is found, it is subtracted from both **X** and **y** and the procedure is re-iterated until **X** becomes a null matrix.

The number of latent variables to be retained in the model can be selected according to different tools. In cross-validation (Wold, 1975), the training data set is split into a number

of subsets, say *r*. Initially, for a model comprising one latent variable, the first subset of data is omitted and a PLS model is built on the remaining (*r-1*) subsets of data. The prediction error sum of squares (PRESS) for the omitted subset of data is then computed and the omitted subset restored. The procedure is repeated until every individual subset has been left out once. The *r* individual PRESS's are then summed to give the total PRESS. The procedure is repeated for $i=\{2,3,\ldots, a\}$ latent variables and a corresponding total PRESS is calculated. The optimal number of latent variables is chosen to be that which minimizes the total PRESS. A nice description of PLS Regression can be found in Tenenhaus (1998) and Camminatiello (2006).

3. The review of fuzzy linear regression

Since the first FLR model proposed by Tanaka (1982), the fast growing literature has followed two paths. One of which merely depends on developing new parameter estimation methods which enable to compute less fuzzier and more useful parameter estimates. Some of them mentioned are given in Soliman et.al. (2002), Toyoura et. al. (2004), Chang (2001), Alex (2006), Ishibuchi and Nii (2001), Tran and Duckstein (2002). In general, these methods can be categorized into two classes, which are called mathematical programming based parameter estimation methods and fuzzy least squares method, respectively. Former ones are those that are based on mathematical programming. Later ones are based on the method proposed by Diamond (1988). Both aim to improve parameter estimates. Second of which is based on application of the model. However, this aspect of FLR has not got much attention. Generally speaking, the issues emerging from applications such as modelling several explanatory variables, interactions among them have been avoided. FLR is a method which is more suitable when one or more of the violations occur simultaneously, for example, the assumption of linearity between dependent and independent variables may not be observed, or instead of numeric data values, data related to one or more variables can be described as words such as "*bad*" or "*good*" or there exists small data set which does not satisfy the normality assumption. Under these circumstances, classic regression is observed to fail. This situation is exemplified with a solid example in the next section.

Also, Kim et. al. (1996) investigated various circumstances where classic regression excels fuzzy regression or vice versa.

Fuzzy linear regression model is generally given as follows:

$$\hat{\tilde{Y}}_i = \tilde{A}_0 + \tilde{A}_1\tilde{X}_1 + \tilde{A}_2\tilde{X}_2 + \ldots + \tilde{A}_n\tilde{X}_n \qquad (3.1)$$

where $\hat{\tilde{Y}}_i, \tilde{X}_i, \tilde{A}_j$ denotes fuzzy numbers which can be symmetric or asymmetric or trapezoidal fuzzy numbers.

Symmetric or asymmetric or trapezoidal fuzzy numbers can be chosen based on information which will be believed that it represents inherent uncertainty in FLR. For example, it is believed that asymmetric fuzzy numbers represent uncertainty in parameters. Then, the model

is constructed based on those numbers. As it can been seen, the expression in (3.1) exhibits the generic case for FLR. The model given in (3.1) does not have error term since it is included in the parameters of model. The special forms of model (3.1) can be written

depending on the type of variables. The Table 1 summarizes the cases which should be used in modeling.

Table 1: *Type of variables*

Y	X	A
Reel	Reel	Fuzzy
Fuzzy	Reel	Fuzzy
Fuzzy	Fuzzy	Fuzzy

Throughout the paper symmetric triangular fuzzy numbers are employed for the sake of simplicity. Linear programming based method is used in order to estimate parameters.

4. Implementing fuzzy linear regression

The parameter estimation method used in this paper is based on the method proposed by Tanaka (1982). For this purpose, the formula below is employed to estimate parameters of FLR.

$$
\begin{aligned}
&Min\ \mathbf{c}^t |\mathbf{X}| = Min \sum_{j=0}^{n} c_j \sum_{i=1}^{m} |x_{ij}| \\
&\sum_{j=0}^{n} \alpha_j x_{ij} + (1-h)\sum_{j=0}^{n} c_j |x_{ij}| \geq y_i + (1-h)e_i \\
&\sum_{j=0}^{n} \alpha_j x_{ij} - (1-h)\sum_{j=0}^{n} c_j |x_{ij}| \leq y_i - (1-h)e_i \\
&c_j \geq 0,\ \alpha \in \Re,\ x_{i0} = 1\ (0 \leq h \leq 1;\ \forall i = 1,2,...,m)
\end{aligned}
\qquad (4.1)
$$

To illustrate why classic regression failed, we used a data set (Tanaka and Guo, 1999) whose name is *House price* which is given in Table 2.

Table 2: *house price data*

N	y	x_1	x_2	x_3	x_4	x_5
1	606	1	38,09	36,43	5	1
2	710	1	62,1	25,5	6	1
3	808	1	63,76	44,71	7	1
4	826	1	74,52	38,09	8	1
5	865	1	75,38	41,1	7	2
6	852	2	52,99	26,49	4	2
7	917	2	62,93	26,49	5	2
8	1031	2	72,04	33,12	6	3
9	1092	2	76,12	42,64	7	2
10	1203	2	90,26	43,06	7	2
11	1394	3	85,7	31,33	6	3
12	1420	3	95,27	27,64	6	3
13	1601	3	105,98	27,64	6	3
14	1632	3	79,25	66,81	6	3
15	1699	3	120,5	32,25	6	3

The explanatory variables $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$and $\mathbf{x}_5$ are quality of the construction material, area of the first floor, area of the second floor, total number of rooms, number of Japanese room, respectively. The response variable **y** is the price of houses whose last four digits are dropped for the sake of simplicity.

When classic regression analysis is used, the model is obtained as follows:

$$\mathbf{y}=-112.4+236.48\mathbf{x}_1+9.3568\ \mathbf{x}_2+8.2294\ \mathbf{x}_3-37.889\ \mathbf{x}_4-17.253\ \mathbf{x}_5 \quad (4.2)$$

It is observed that as the x_4 (total number of rooms) increases**, y** (price) decreases. This is contradictory to common sense. Therefore, FLR can be substituted as an alternative modeling approach. However, this also creates problems which should be addresses too. The same data produce FLR as follows:

$$\hat{\tilde{y}} = (45.167,37.634)x_1 + (5.833,0)x_2 + (4.786,0)x_3 \quad (4.3)$$

Also, this model explains response variable using fewer variables although there is no procedure available in FLR which can be used as variable selection method. Therefore, when several explanatory variables exist in FLR, it should be expected that some problems similar to those in regression or the problems related to FLR may emerge. To overcome these kinds of problem, PLS is an alternative method that can also be used in FLR. In order to illustrate the usage of PLS in FLR, a data set consisting of seven variables given in Table 3 are used. Based on the results of PLS, just one variable (component) which can be written in the form of other variables is obtained as combination of other variables. Then FLR is conducted for this variable.

5. Implementing PLS into fuzzy linear regression

On consider the following data sample. The data consisting of the price, weight and nutritional information was gathered for a number of chocolates commonly available in Queensland stores. The data was gathered in April 2002 in Brisbane. There are 7 varieties and 7 variables, plus the names of the chocolates are row names**.**

Table 3: *Chocolates Data*

N	Unit.Price	Size	Energy	Protein	Fat	Carbohydrate	Sodium
1	1,76	50	1970	3,1	27,2	53,2	75
2	2,56	45	2250	7,2	30,1	59,4	110
3	1,62	60	1890	4,7	19,5	67,9	160
4	2,56	50	2030	5,6	20,4	67,4	250
5	2,33	55	1623	2,2	9,2	73,3	90
6	2,58	60	1980	8,5	20,6	63,3	130
7	2,78	42.5	1970	5	20	69	148

As it can be seen in the previous section, FLR may fail if some of the explanatory variables are correlated when independent and dependent variables are crisp. This situation also creates problems for FLR. To overcome this kind of a problem, a method called PLS used

in classic regression can also be used in FLR. The Chocolate Data consists of six independent and one dependent variable. The dependent variable price is tried to be explained by independent variables such as size, energy and so on. In this example, the classic regression failed. Then FLR is run to estimate parameters but the similar situation observed in the previous example is encountered. One of the frequently faced problems in classic regression has appeared. It is called correlated explanatory variables. This problem is expected since independent and dependent variables are crisp values. Then PLS is used to construct new variables (components) to be used in FLR model. For our case, After running PLS, one independent variable (component) is obtained which is denoted by X*.
When the linear programming formula is run for the data obtained after PLS, the resulting fuzzy regression model is obtained as follows:

$$\hat{y} = (2.14,0.78) + (4.27,1.12)X^{*} \tag{5.1}$$

where X* is the component which is a linear combination of the independent variables after PLS is calculated.

The constant term of FLR is (2.14, 0.78) and the coefficient of X is (4.27, 1.12). These are symmetric fuzzy numbers which can be written as (1.36, 2.92) and (3.15, 5.39). Suppose that X*=0.25, then the predicted price is (2.14, 0.78)+(4.27, 1.12)0.25= (3.21, 1.06) is obtained. This means that the price ranges between 2.15 and 4.27 when X*=0.25.

Instead of using correlated explanatory variables, the component, which is a linear combination of six independent variables such as size, energy, protein, fat, carbohydrate, and sodium, is used to estimate the price by using FLR.

6. Conclusion

When the assumptions related to classic regression are violated such as correlated independent variables or correlated errors or other types of violations that can be found in the literature, some type of remedies are suggested. When functional relationship is not known in advance, FLR is introduced as an alternative method which helps model crisp/crisp or crisp/fuzzy data. On the other hand, PLS is used for reducing number of independent variables to obtain components. What it is observed in classic regression as problems is also observed in FLR as well. Thus, the methods used for regression can be used for FLR. In this paper, we use PLS for FLR. In the first data set which is house data set, first of all, parameter estimates are calculated for regression model but there is contradiction between one of the independent variable and dependent variable which is that when the number of rooms increase, the price of house decreases. Also, the correct functional relationship between dependent variable and independent variable is not know. Then, the parameters of FRL is calculated. This model has three independent variables. However, FLR is more realistic than classic regression. In the second data set which is called Chocolate Data, the similar problem is encountered. We followed the similar steps to reach the final regression model since the relationship between the dependent variable and the interdependent variables are unknown. Therefore, this led to choose the FLR as an alternative modeling tool. Before running FLR, PLS is employed to reduce the number of independent variables. Then based on the reduced number of independent variable, which is one component consisting of

linear combination of independent variables. Then we predicted FLR model using component as an independent variable and the price as a dependent variable. Therefore, what model suggested is that the component variable as a combination of independent variables explains the price in the interval. As a further study, extending PLS method to be used for the case of fuzzy/fuzzy is a subject which should be examined.

References

1. Alex, R., **A new kind of fuzzy regression modelling and its combination with fuzzy inference**, 10, Soft Computing, 2006, pp. 618-622.

2. Camminatiello I., **Robust methods for Partial Least Squares Regression: methodological contributions and applications in environmental field**. PhD thesis, University of Naples Federico II, 2006.

3. Chang, Y.H., **Hybrid fuzzy least-squares regression analysis and its reliability measures,** Fuzzy Sets and Systems,119, 2001, 225-246.

4. Diamond, P., **Fuzzy least squares**, Information Science ,46, 1988, pp. 141-157.

5. Garthwaite., P.H., **An Interpretation of Partial Least Squares**, Journal of the American Statistical Association, Vol. 89, No. 425., 1994, pp. 122-127.

6. Ishibuchi, H., Nii, M., **Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks**, Fuzzy Sets and Systems, 119, 2001, pp. 273-290.

7. Kim, K.J, Herbert, M., Koksalan, M, **Fuzzy versus statistical linear regression**, European Journal of Operational Research, 92, 1996, pp. 417-434.

8. Soliman, S.A., Alammari, R.A., Temraz, H.K., El-Hawary, M.E., (2002), Fuzzy linear parameter estimation algorithms: a new formulation,Electrical Power and Energy Systems, 24, pp. 415-420.

9. Tenenhaus M., **La Rgression PLS, Thorie et Pratique**. Editions Technip, Paris, 1998.

[10] Toyoura, Y., Watada, J., Khalid, M., Yusof, R., **Formulation of linguistic regression model based on natural words**, Soft Computing,8, 2004, pp. 681-688.

[11] Tanaka, H., Uejima, S., Asai, K., **Regression analysis with fuzzy model, IEEE Transactions on Systems**, Man, and Cybernetics SMC-12, 1992, pp. 903-907.

[12] Tanaka,H., Guo, P., **Possibilistic Data Aanalysisi for Operation Reserach**, Springer-Verlag, New York, 1999.

[13] Tran, L., Duckstein, L., **Multiobjective fuzzy regression with central tendency and possibilistic properties**, Fuzzy Sets and Systems,130, 2002, 21-31.

[14] Valle, S., W. Li, and S. J. Qin, **Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods**. Ind. Eng. Chem. Res., 38, 1999, pp. 4389-4401.

[15] Wold, H., **Soft modelling by latent variables: Non linear Iterative Partila Least Squares approach. In Perspectives in Probability and Statistics: Papers in honour of Bartelett**, J. Gani (ed.), 1975, pp. 117-142. London: Ac. Press,

[16] Zadeh, L.A., **Fuzzy sets as a basis for a theoryof possibility**, Fuzzy Sets and Systems, 1(1), 1978 pp. 3-28.

---

[1]Alper Basaran is assistant professor at Department of Mathematics, Nigde University (Turkey) from 2008. He gained the B.Sc. in Statistics at Hacettepe University in 1993, which is one of the most competitive universities in Turkey. Then He was supported by The Minsitry of Education in Turkey with a scholarship for studying in the M.Sc. in Mathematics in U.S.A from 1997-2000. He completed the Ph.D in Statistics at Hacettepe University in Turkey in 2007.

[2] Biagio Simonetti is a researcher in Statistics at the Università degli Studi del Sannio and also a professor in Statistics and Statistics for Business. He received his PhD in Computational Statistics. His research interests focus on Multivariate Statistical Methods with particular attention to Correspondence Analysis applied to problem of the evaluation of the Customer Satisfaction. He serves on the program and organizing committees of conferences in the field of Multivariate Methods and Decision Theory. He is author of more than 50 conference and high impact factor journal papers as Journal of Multivariate Analysis, Journal of Applied Statistics and Communication in Statistics – Theory and Methods.

[3] Luigi D'Ambra is full professor of Statistics at University of Naples Federico II (Italy). He has carried out an intense didactic and scientific activity. He has supervised numerous theses of statistics on the topic of the customer satisfaction in the context of Public services (transport, education, health). In the last years he has been taken coordinator of national project research for the Customer Satisfaction in Health Services. He held seminaries, lessons about multivariate analysis, Partial Least Squares, and the multidimensional scaling, with applications to the public services: education and health services in the context of the PhD of Computational Statistics in many Italian University.