FUZZY REGRESSION METHODOLOGY

Prajneshu
I.A.S.R.I., Library Avenue, New Delhi – 110 012
prajnesh@iasri.res.in

1. Introduction
It is, by now, well recognized that Agriculture is a “Soft science”, unlike Physics or Chemistry, which are “Hard sciences”. In the former, there is always some amount of “impreciseness” or “vagueness” or “fuzziness” in the underlying phenomenon, and/or explanatory variables, and/or response variable(s). Therefore, for a more realistic modelling, there is a need to incorporate this aspect in traditional models, like Multiple linear regression model.

2. Fuzzy Linear Regression Methodology
In conventional regression analysis, deviations between observed and estimated values are assumed to be due to random errors. However, quite often these are due to indefiniteness of structure of a system or imprecise observations. Thus, uncertainty in this type of regression model becomes “fuzziness” and not randomness. Studies dealing with Fuzzy linear regression (FLR) model can be broadly classified into two approaches, viz. (i) Linear programming (LP)-based methods, and (ii) Fuzzy least squares (FLS) methods. In the former approach, proposed by Tanaka et al. (1982), parameters of FLR model:

\[ Y = A_0 + A_1 X_1 + \ldots + A_p X_p \]  \hspace{1cm} (1)

where \( A_i = (a_{i0}, a_{iw}) \), \( Y = (y_o, y_w) \) are estimated by minimizing “Total vagueness” of model-data combination, subject to constraints that each data point must lie within estimated value of response variable. This can be visualized as a LP problem and solved by using “Simplex procedure”. Several software packages, like SAS, LP88, and LINDO are available for solving LP problem. Any standard spreadsheet package, like Microsoft Excel may also be used to solve LP problem manually.

Kandala and Prajneshu (2003) demonstrated applicability of above methodology when the two explanatory variables (viz. Plant height and Leaf area index) and response variable (Dry-matter accumulation) are all crisp but underlying phenomenon is assumed to be fuzzy in nature. It was shown that widths of prediction intervals in respect of Fuzzy linear regression model were much less than those for Multiple linear regression model. Kandala and Prajneshu (2002) had earlier obtained similar results in the situation when the two explanatory variables, viz. Normalized difference vegetation index (NDVI) and Ratio vegetation index (RVI) are highly correlated.

Further, for determining age-length relationship in a fish species, response variable (length) generally lies in an interval for different fish of same age. Kandala and Prajneshu (2004) applied FLR methodology for fitting fuzzy von Bertalanffy growth model with a view to determining age-length relationship in pearl oyster. It may be pointed out that traditional statistical methods are not capable of handling such a situation in which response variable is in intervals. The only way out there is to get rid of interval values for response variable to
crisp values either by taking mean or mode, thereby losing a lot of vital information about spread.

However, a criticism of Tanaka’s approach is that it is not based on sound statistical principles. Another drawback, as pointed out by Chang and Ayyub (2001) and D’Urso (2003), is that as the number of data points increases, the number of constraints in LP increases proportionally, thereby resulting in computational difficulties.

The second approach based on Fuzzy least squares (FLS) method, was pioneered by Diamond (1988), which as its name suggests, is a fuzzy extension of Least squares method based on a new defined distance on the space of fuzzy numbers. Kandala and Prajneshu (2004) have applied this methodology for fitting well-known “Allometric model” to length-weight data of some fish species. However, a drawback of this procedure is that the spread of estimated responses increases as magnitude of explanatory variable increases, even though the spread of observed responses are roughly constant or decreasing. To overcome this, Kao and Chyu (2002) proposed a “two-stage” approach for fitting FLR model through FLS approach and showed its superiority over Diamond’s procedure. Recently, Singh et al. (2007) have thoroughly discussed this approach and, for its application, relevant computer programs have been developed in “Nonlinear programming solver LINGO, Version 8” software package. Finally, fitting of fuzzy von Bertalanffy growth model, when response variable is reported in intervals corresponding to various crisp values of explanatory variable, is carried out for pearl oyster age-length data. Yang and Liu (2003) developed robust algorithms against presence of outliers in a FLR model. However, this methodology still needs to be applied to data from the field of agriculture.

3. Fuzzy Nonlinear Regression Methodology

In contrast to fuzzy linear models, very little research work so far has been done dealing with “Fuzzy nonlinear models (FNMs)”. Buckley and Feuring (2000) proposed “Evolutionary algorithm solutions” for fitting some particular parametric FNMs. Specifically, for given fuzzy data, the algorithm searches from the “Library” of fuzzy functions (which includes linear, polynomial, exponential, and logarithmic) that function which best fits the data. Evidently, this is not at all satisfactory for fitting parametric FNMs to data and so is a fertile possible area for future research.

Hong and Hwang (2003) introduced the use of “Support vector machines” (SVM) for fitting nonparametric FNMs. SVM was developed at AT&T Bell Laboratories by V. Vapnik and his coworkers (Vapnik, 2000) and is based on the idea of Structural risk minimization. An excellent description of various aspects of SVM is given in Suykens et al. (2002). However, a lot of effort still needs to be expanded to fine-tune the methodology in order to apply it to real data.
References


