ELSEVIER

European Journal of Operational Research 92 (1996) 417–434

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Theory and Methodology

Fuzzy versus statistical linear regression [1]

Kwang Jae Kim [a], Herbert Moskowitz [b,*], Murat Koksalan [c]

[a] *Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA*
[b] *Krannert Graduate School of Management, Purdue University, West Lafayette, IN 47907, USA*
[c] *Industrial Engineering Department, Middle East Technical University, Ankara 06531, Turkey*

Received March 1994; revised November 1994

Abstract

Statistical linear regression and fuzzy linear regression have been developed from different perspectives, and thus there exist several conceptual and methodological differences between the two approaches. The characteristics of both methods, in terms of basic assumptions, parameter estimation, and application are described and contrasted. Their descriptive and predictive capabilities are also compared via a simulation experiment to identify the conditions under which one outperforms the other. It turns out that statistical linear regression is superior to fuzzy linear regression in terms of predictive capability, whereas their comparative descriptive performance depends on various factors associated with the data set (size, quality) and proper specificity of the model (aptness of the model, heteroscedasticity, autocorrelation, nonrandomness of error terms). Specifically, fuzzy linear regression performance becomes relatively better, vis-à-vis statistical linear regression, as the size of the data set diminishes and the aptness of the regression model deteriorates. Fuzzy linear regression may thus be used as a viable alternative to statistical linear regression in estimating regression parameters when the data set is insufficient to support statistical regression analysis and/or the aptness of the regression model is poor (e.g., due to vague relationship among variables and poor model specification).

Keywords: Fuzzy regression; Statistical regression; Description; Prediction

1. Introduction

Zadeh (1965) proposed the concept of a fuzzy set which is useful in dealing with classes of problems where there is no sharp transition from membership to nonmembership. Bellman and Zadeh (1970) noted that there is a need for differentiation between randomness and fuzziness. They argued that the premise of 'imprecision can be equated with randomness' is questionable, and that fuzziness is a major source of imprecision in many decision processes.

Statistical linear regression (hereafter, will be referred to as 'statistical regression') analysis is a widely used statistical tool to model the relationship among variables to describe and/or predict

* Corresponding author.

[1] This research was supported in part by the Purdue University Krannert School of Management's Center for the Management of Manufacturing Enterprises (CMME), the Purdue University Engineering Research Center for Intelligent Manufacturing Systems under NSF Grant CDR 850002, and NSF Grant INT 9114147.

SSDI 0377-2217(94)00352-1

phenomena, and is useful in a non-fuzzy environment where the relationship among variables is sharply defined. Fuzzy linear regression (hereafter, will be referred to as 'fuzzy regression'), developed by Tanaka, Uejima and Asai (1982), aims to model vague and imprecise phenomena using the fuzzy functions defined by Zadeh's extension principle (Zadeh, 1975), which provides a general method for extending non-fuzzy mathematical concepts to deal with fuzzy quantities. Fuzzy regression is a nonstatistical method – the deviations between observed and estimated values are assumed to depend on the indefiniteness/vagueness of the parameters which govern the system structure, not on its measurement errors. In this paper, fuzzy regression and statistical regression will be compared both conceptually and empirically.

Fuzzy regression gives rise to a *possibility* distribution that accounts for the imprecise nature or vagueness of our understanding of a phenomenon, which is manifested by yielding fuzzy parameters of the model. Classical statistical regression makes rigid assumptions about the statistical properties of the model; e.g. the normality of error terms and predictions. These assumptions, as well as the aptness of the linear regression model, are difficult to justify unless a sufficiently large data set is available. The violation of such basic assumptions could adversely affect the validity and performance of statistical regression. It has been stated that fuzzy regression may be more effective than statistical regression (or other tools based on statistical techniques) when the latter's assumptions are either violated or cannot be properly employed, as for example, when human judgments are involved (Tanaka, Uejima and Asai, 1982), ambiguous processes must be explained or its outputs must be predicted (Gharpuray et al., 1986; Heshmaty and Kandel, 1985), and when only a small amount of crisp or imprecise data is available (Bardossy, Bogardi and Kelly, 1987; Bardossy, Bogardi and Duckstein, 1990; Gharpuray et al., 1986).

Fuzzy regression can thus be a useful tool for managers and researchers in many decision domains in estimating relationships among variables with fuzzy, incomplete information. For example, in quality function deployment, designers must incorporate both qualitative and quantitative information regarding relationships among customer attributes (e.g., system reliability) and engineering characteristics (e.g., door seal resistance) into the problem formulation (Kim et al., 1995).

There thus is a need to investigate the comparative characteristics and performance of statistical versus fuzzy regression. Our purpose is to describe the conceptual and methodological differences between the two methods, and then to report the results of a simulation experiment, which identifies the conditions under which one approach outperforms the other by comparing their descriptive and predictive capabilities. This experiment considers such factors as the size and quality of the data set, the correctness of the model specification, the presence of heteroscedasticity, autocorrelation, and nonrandomness of the error terms.

In Section 2, the conceptual and technical differences between statistical and fuzzy regression, as well as their characteristics are described. The simulation experiment and analysis of the results are discussed in Section 3. The implications of our findings are presented in Section 4.

2. Comparison of characteristics

Statistical regression and fuzzy regression have been developed from different perspectives and constructs, and thus there exist major differences between the two approaches; for example, in their basic assumptions, method of parameter estimation, and usage. The following subsections compare the two approaches from a conceptual and methodological perspective.

2.1. Basic assumptions

In statistical regression, the deviation of the observed value from the estimated value, viz. residual, is assumed to derive from the following two sources: 1) relevant factors omitted explicitly

from the model, and 2) random measurement errors in recording observations. Further, it is generally assumed that statistically the error terms are normally distributed. Fuzzy regression, on the other hand, attributes the error to the indefiniteness of the system structure and the vagueness of human perception or knowledge of the model (Tanaka, Uejima and Asai, 1982; Gharpuray et al., 1986). Such ambiguity is reflected in the regression model via fuzzy parameters.

The difference in the basic assumptions of the two methods is essentially due to their individual view of and approach to uncertainty. Statistical regression is based upon classical set theory, and employs the concepts and techniques of probability theory, in which imprecision is synonymous with randomness. Fuzzy regression is based on fuzzy set theory. Hence, the degree of membership of an object to a set is defined by a membership function, and possibility measures provide a basis for the manipulation of the involved fuzzy sets.

2.2. Parameter estimation

2.2.1. Statistical regression

A basic regression model where there are n independent variables and the regression function is linear can be stated as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \varepsilon_i, \tag{2.1}$$

where y_i is the value of the response variable in the i-th observation, x_{ij} is the value of the j-th independent variable in the i-th observation ($j = 1, \ldots, n$), β_j's are parameters ($j = 0, 1, \ldots, n$), and ε_i is a random error term. The error term ε_i is assumed to follow a distribution with mean zero and constant variance for all i, and covariance

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \text{for all } i, j \quad i \neq j.$$

Based on these assumptions, the regression parameters can be estimated. If the normality of the error term is assumed, the values of the dependent variable can be predicted for given values of independent variables. Confidence intervals can be easily constructed for regression parameters and the mean and individual values of the dependent variable.

2.2.2. Fuzzy regression

Consider a fuzzy linear function

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = AX,$$

where $A = (a_0, a_1, a_2, \ldots, a_n)$ are unknown fuzzy parameters, which can be denoted in vector form as

$$A = \{\alpha, c\},$$

$$\alpha = (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n),$$

$$c = (c_0, c_1, c_2, \ldots, c_n).$$

Here, α_j is the mean or center value of a_j, and c_j is its dispersion or spread, $j = 0, \ldots, n$. Letting

$$X_i = (1, x_{i1}, x_{i2}, \ldots, x_{in}),$$

the problem in the fuzzy linear regression model is to determine fuzzy parameters A^* such that the membership value of y_i (the i-th observed value of the dependent variable) to its fuzzy estimate $y_i^* = A^* X_i$ is at least H, which is selected by a decision maker (DM), for all i, $i = 1, \ldots, N$. The H value, which is between 0 and 1, is referred to as the degree of fit of the estimated fuzzy linear model to the given data set. A physical interpretation of H is that y_i is contained in the support interval of y_i^* which has a degree of membership $\geq H$, for all i (Tanaka, Uejima and Asai, 1982). This condition can be represented as a pair of inequality constraints for each set of observations i as follows:

$$\alpha_0 + \sum_{j=1}^{n} \alpha_j x_{ij} + |L^{-1}(H)| \left(c_0 + \sum_{j=1}^{n} c_j |x_{ij}| \right) \geq y_i. \tag{2.2}$$

$$\alpha_0 + \sum_{j=1}^{n} \alpha_j x_{ij} - |L^{-1}(H)| \left(c_0 + \sum_{j=1}^{n} c_j |x_{ij}| \right) \leq y_i. \tag{2.3}$$

where L represents the reference or membership function of a standardized fuzzy parameter, i.e.,

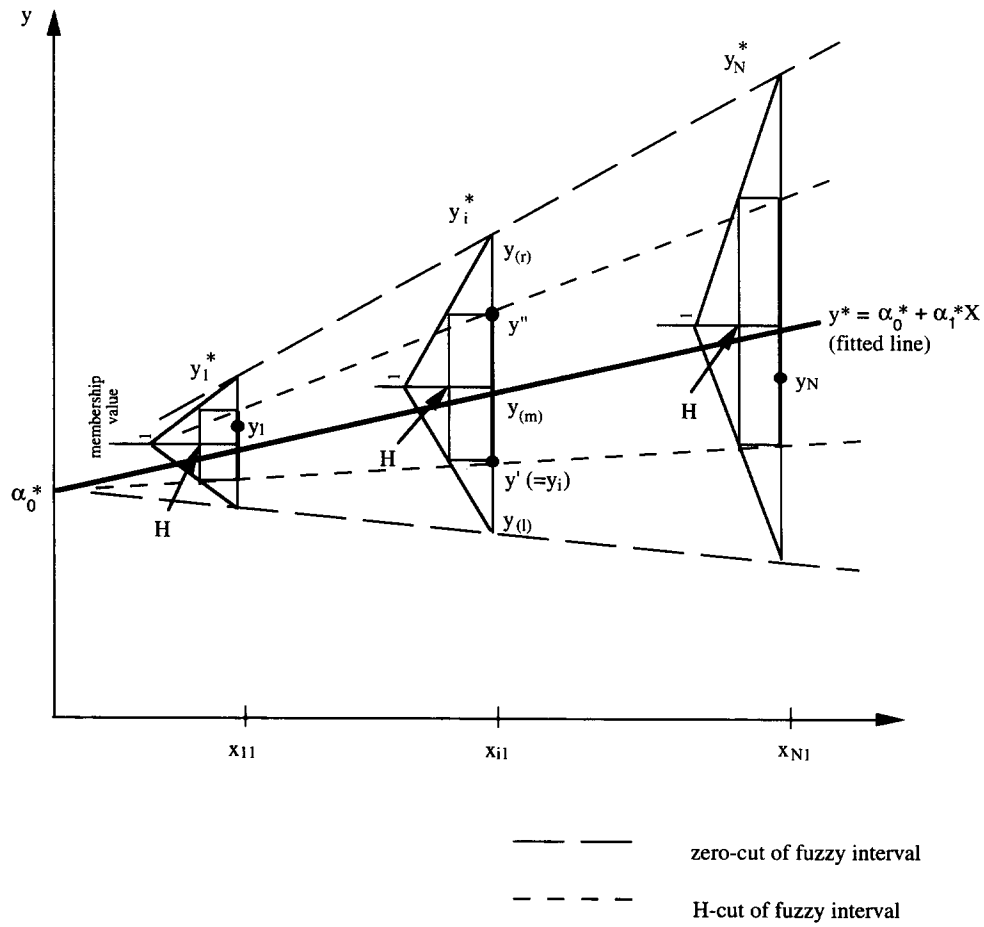


Fig. 1. Interpretation of fuzzy regression algorithm.

distance from the center value measured in units of spread (Tanaka and Watada, 1988). [1]

$$L(x) = \max(0, 1 - |x|^p), \quad p > 0,$$

is an illustrative membership function (Dubois and Prade, 1980). Our aim is to minimize the fuzziness in the predicted value of dependent variable y_i. This can be achieved by minimizing the sum of spreads of all fuzzy parameters a_j, i.e., minimizing $(c_0 + c_1 + c_2 + \cdots + c_n)$, which is analogous to the least squares criterion in statistical regression analysis, under the restriction of (2.2) and (2.3). [2] The fuzzy linear regression problem can be posed as an equivalent linear program as follows (Tanaka, Uejima and Asai, 1982):

Find α and c which

$$\text{Minimize} \quad c_0 + c_1 + c_2 + \cdots + c_n \tag{2.4}$$

subject to

$$\text{(2.2) and (2.3)}, \quad i = 1, \ldots, N, \tag{2.5}$$

$$c_j \geq 0, \quad j = 0, 1, \ldots, n. \tag{2.6}$$

[1] Eqs. (2.2) and (2.3) correspond to a symmetric fuzzy parameter or symmetric membership function, i.e., $L = R$ in L-R representation (Dubois and Prade, 1980). The equations can also be modified for an asymmetric membership function (Bardossy, 1990).

[2] Different forms of indices of fuzziness are proposed in Tanaka (1987) and Bardossy (1990).

Fig. 1 is a visualization of the constraints of the linear program given above when a linear membership function is employed, and the model has one independent variable and an intercept term. In Fig. 1, $y_{(m)}$ corresponds to $(x_{i1}, \alpha_0 + \alpha_1 x_{i1})$, and the points y', y'', $y_{(1)}$ and $y_{(r)}$ correspond to

$$(x_{i1}, (\alpha_0 + \alpha_1 x_{i1}) - |L^{-1}(H)|(c_0 + c_1 x_{i1})),$$

$$(x_I, (\alpha_0 + \alpha_1 x_{i1}) + |L^{-1}(H)|(c_0 + c_1 x_{i1})),$$

$$(x_{i1}, (\alpha_0 + \alpha_1 x_{i1}) - (c_0 + c_1 x_{i1})),$$

and

$$(x_{i1}, (\alpha_0 + \alpha_1 x_{i1}) + (c_0 + c_1 x_{i1})),$$

respectively, where x_{i1} is the value of the independent variable at the i-th observation. The interval $[y', y'']$ corresponds to the H-cut of y_i^*; namely, any point in $[y', y'']$ yields a membership value of at least H to y_i^*. Thus, if the i-th observation y_i falls in $[y', y'']$, the condition represented in (2.2)–(2.3) is satisfied.

Because the degree of membership of y_i to its fuzzy estimate y_i^* increases as the center value of y_i^* gets closer to y_i, the H parameter can be viewed as denoting the level of credibility or confidence that the DM desires; i.e., for a given degree of credibility (viz. H), the fuzzy regression algorithm determines the spreads (and the center values as well) of the parameters to satisfy the H level specified. Accordingly, a higher level of credibility yields a wider spread, just as a higher confidence coefficient in statistical regression yields a wider confidence interval. For example, 100% credibility (i.e., $H = 1$) gives an extremely wide range on the fuzzy regression parameters. Conversely, for a given spread on the parameters, H indicates the highest confidence that is attainable. (The H value, membership function shape, and spreads of fuzzy parameters are closely interrelated. A proper H value which incorporates a DM's beliefs regarding the shape and range of the possibility distribution involved can be assessed systematically (Moskowitz and Kim, 1993).)

The distance between y' and $y_{(m)}$ or between $y_{(m)}$ and y'' is

$$|L^{-1}(H)|(c_0 + c_1 x_{i1}),$$

which denotes the maximum deviation allowed between the center value of y_i^* ($y_{(m)}$) and the observation y_i to maintain the level of credibility of at least H (Fig. 1). Because the i-th observation y_i coincides with the left extreme of the H-cut, $[y', y'']$, the level of credibility achieved at this observation is equal to H. A more general case where the dependent variable (y_i) is also fuzzy can be easily accommodated in a similar fashion (Tanaka, Uejima and Asai, 1982).

2.3. Usage

Both statistical and fuzzy regression first require that the model be specified (i.e., selection of independent variables and choice of the functional form of the regression equation) before using the data set to estimate its parameters. Fuzzy regression additionally requires prespecifying the H value (or credibility level).

The parameter estimates of statistical regression are crisp numbers, while those of fuzzy regression are fuzzy, consisting of a spread as well as center value. The degree of membership of any real number in the interval [center value − spread, center value + spread] is defined by an associated membership function. Because the parameters in fuzzy regression are fuzzy numbers, the estimated dependent variable is also a fuzzy number, yielding a fuzzy interval of the dependent variable. Tanaka and Watada (1988) show that the center value is not affected by the H parameter value, however, a higher H value induces a larger spread. Moreover, as the membership function becomes decreasingly concave or increasingly convex, the spread enlarges (Moskowitz and Kim, 1993).

With respect to fuzzy regression, after model specification and parameter estimation are completed, there is no provision for inferences or tests of the appropriateness of the model, since it is a nonstatistical method. However, in statistical regression, various statistical analyses can be invoked to provide more information about the properties of the model – inferences concerning model parameters, interval estimation of the mean responses, prediction of new observations, analysis of variance for goodness of fit, etc. Here, the DM specifies a confidence level (e.g., 95%),

which is analogous to the H parameter value in fuzzy regression, to indicate the degree of confidence placed on the interval estimate or hypothesis test.

The difference between the confidence interval in statistical regression and the fuzzy interval in fuzzy regression is worth mentioning. Consider, for example, a 95% confidence interval for the mean response in statistical regression and a fuzzy interval for the dependent variable corresponding to $H = 0.95$ in fuzzy regression. The coefficient 0.95 of the confidence interval implies that if many independent samples are taken at the same levels of the independent variables, and a 95% confidence interval is constructed for each sample, then 95% of the intervals will contain the true value of the mean of the dependent variable. Therefore, the confidence interval or confidence coefficient should be interpreted in relation to repeated sampling or future predictions. However, a fuzzy interval corresponding to $H = 0.95$ simply denotes the narrowest interval whose 0.95-cut includes every observation in the given sample. In other words, every observation has a membership value of at least 0.95 to its fuzzy estimate defined by the fuzzy interval. For instance, the interval $[y', y'']$ in Fig. 1 is a fuzzy interval with $H = 0.95$ at $x = x_{i1}$ if H is set at 0.95. Thus, the focus of the fuzzy interval is entirely on the given observations, not on future sampling or predictions. Therefore, unlike statistical regression, fuzzy regression (or a fuzzy interval), per se, is not theoretically or conceptually appropriate for making predictions.

2.4. Distribution of the dependent variable

In statistical regression, the error terms (ε_i) are generally assumed to be independent and normally distributed with mean zero and constant variance σ^2 ($N(0, \sigma^2)$). Since the dependent variable, y_i, is a linear function of ε_i (see (2.1)), it is well known that the y_i's are independent:

$$N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}, \sigma^2).$$

Thus, the distribution of y_i is symmetric, bell-shaped, and has a constant standard deviation at all x levels. However, in fuzzy regression, the shape of the possibility distribution of y_i is essentially determined by the assumed membership function of the fuzzy model parameters. Fuzzy regression problems in the literature usually employ a linear, symmetric, triangular membership function. Nonlinear membership functions and/or asymmetric membership functions with different shapes for each model parameter can also be employed (Bardossy, 1990). In terms of the shape of the distribution, fuzzy regression therefore has more flexibility than statistical regression.

Without loss of generality, we now compare the sampling distribution of a point estimator ($\hat{y}$) of the mean response ($E(Y)$) in statistical regression and a possibility distribution of the fuzzy estimate (y^*) in fuzzy regression with one independent variable. Both the point estimator ($\hat{y}$) and fuzzy estimate (y^*) are used to provide the estimated mean (or center) values of the dependent variable at a fixed level of the independent variable, say x_h. In addition to the mean values, they also provide measures of the variability of the estimated mean values, namely, standard deviation and spread. The standard deviation of $\hat{y}$ is smallest when $x_h = \bar{x}$, and increases hyperbolically as x_h deviates from $\bar{x}$, where $\bar{x}$ denotes the mean of the independent variable values of the given data set. [3] The spread of the possibility distribution of y^*, however, is smallest when x_h coincides with the origin (i.e., $x_h = 0$), and increases proportionally to the absolute value of x_h.

The magnitude of the standard deviation and the fuzzy spread show different patterns of behavior as the number of sample observations changes. The standard deviation of $\hat{y}$ decreases as more data are available, whereas the spread of y^* increases with more data. The behavior of the standard deviation is well explained using statisti-

[3] The standard deviation of $\hat{y}$ is

$$\left[\mathrm{MSE}\left[\frac{1}{N} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]\right]^{1/2}$$

in simple linear regression, where

$$\mathrm{MSE} = \frac{\sum (y_i - \hat{y}_i)^2}{N - 2}$$

and N is the number of data points.

cal theory; conceptually, a larger sample size provides more (and therefore more accurate) information regarding the population mean response. Namely, the standard deviation is inversely proportional to the square root of the sample size.

The rationale for the opposite behavior of the fuzzy spread can be explained by the concept of possibility. Fuzzy regression determines the spreads of fuzzy parameters (as well as center values) in such a way that every observation in the sample is contained in the H-cut of the corresponding fuzzy estimate. In possibility models, each data point represents a part of the possibility existing in the underlying model. The fuzzy regression algorithm requires that the fitted fuzzy model explain all of the possibilities given in the data set. Thus, having more data points means having more possibilities that should be explained by the fuzzy model. A fuzzy model explains possibility by sacrificing precision, i.e., allowing or increasing spreads. Therefore, the estimated fuzzy spread increases with more data.

3. Comparison of capabilities

The major uses of regression analysis include 1) the description or explanation of a relationship among variables by estimating the model parameters, and 2) the prediction of the dependent variable values given the levels of independent variables. In this section, the descriptive and predictive capabilities of statistical regression vis-à-vis fuzzy regression are compared. A simulation experiment was designed and executed to identify conditions under which each method performs better.

Table 1
Definition of experimental factors

Factor	Definition	Levels	Explanation
N	Number of data points used to estimate the regression parameters	SMALL	Small number of data points
		LARGE	Large number of data points
Q	Quality of data set (measure of dispersion of data points)	GOOD	Error variance is small.
		BAD	Error variance is large.
MA	Aptness of the regression model	GOOD	There is one independent variable, and it is included in the regression model.
		BAD	There are two independent variables, one of which is more important than the other. The more important variable is omitted from the regression model.
HS	Heteroscedasticity of error terms	NONE	Error variance is constant.
		SEVERE	Error variance increases with x severely.
AC	Autocorrelation of error terms	NEGATIVE	Large negative autocorrelation
		ZERO	Zero autocorrelation
		POSITIVE	Large positive autocorrelation
NR	Nonrandomness of error terms	SMALL	Very small number of runs
		MEDIUM	Median number of runs (most random)
		LARGE	Very large number of runs

3.1. Design of experiment

The underlying model considered in this experiment was linear, had two independent variables, and is given in the following equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad (3.1)$$

where y_i is the value of the dependent variable in the i-th observation, x_{ij} is the value of the j-th independent variable in the i-th observation ($j = 1, 2$), β_j's are model parameters ($j = 0, 1, 2$), and ε_i is a random error term which is normally distributed with mean zero.

Two types of factors were considered. The first type was not directly related to the assumptions of statistical or fuzzy regression analysis, and included: 1) the size of the data set (N), 2) the quality of the data set (Q), and 3) the aptness of the regression model (MA). The second type of factors was directly related to the assumptions of statistical regression, and included: 1) heteroscedasticity (HS), 2) autocorrelation (AC), and 3) nonrandomness (NR) of the error terms. The definition of factors and their brief descriptions are given in Table 1 (actual factor values are given in Section 3.3). Factors AC and NR in Table 1 cannot be implemented in the same experiment. [4] Thus, we performed two separate sets of experiment; one set considered factors N, Q, MA, HS, and AC (this experiment will be referred to as NQMHA), while the other set considered factors N, Q, MA, HS, and NR (referred to as NQMHN). Each experiment involved a full factorial design (i.e., $2^4 * 3$). For each treatment, data was generated according to the factor levels, statistical regression and fuzzy regression were then applied to the generated data set, and performance was measured. This process was replicated 20 times.

[4] The order of the data points was randomly rearranged after they had been generated so that the resulting sequence of the data set could have the specified number of runs (factor NR). This rearranging operation and the implementation of the autocorrelation (factor AC) level are not compatible, in the sense that both of them cannot be achieved simultaneously.

3.2. Performance measures

The performance measures developed for descriptive and predictive capabilities are described below.

3.2.1. Measure of descriptive capability: VOLUME

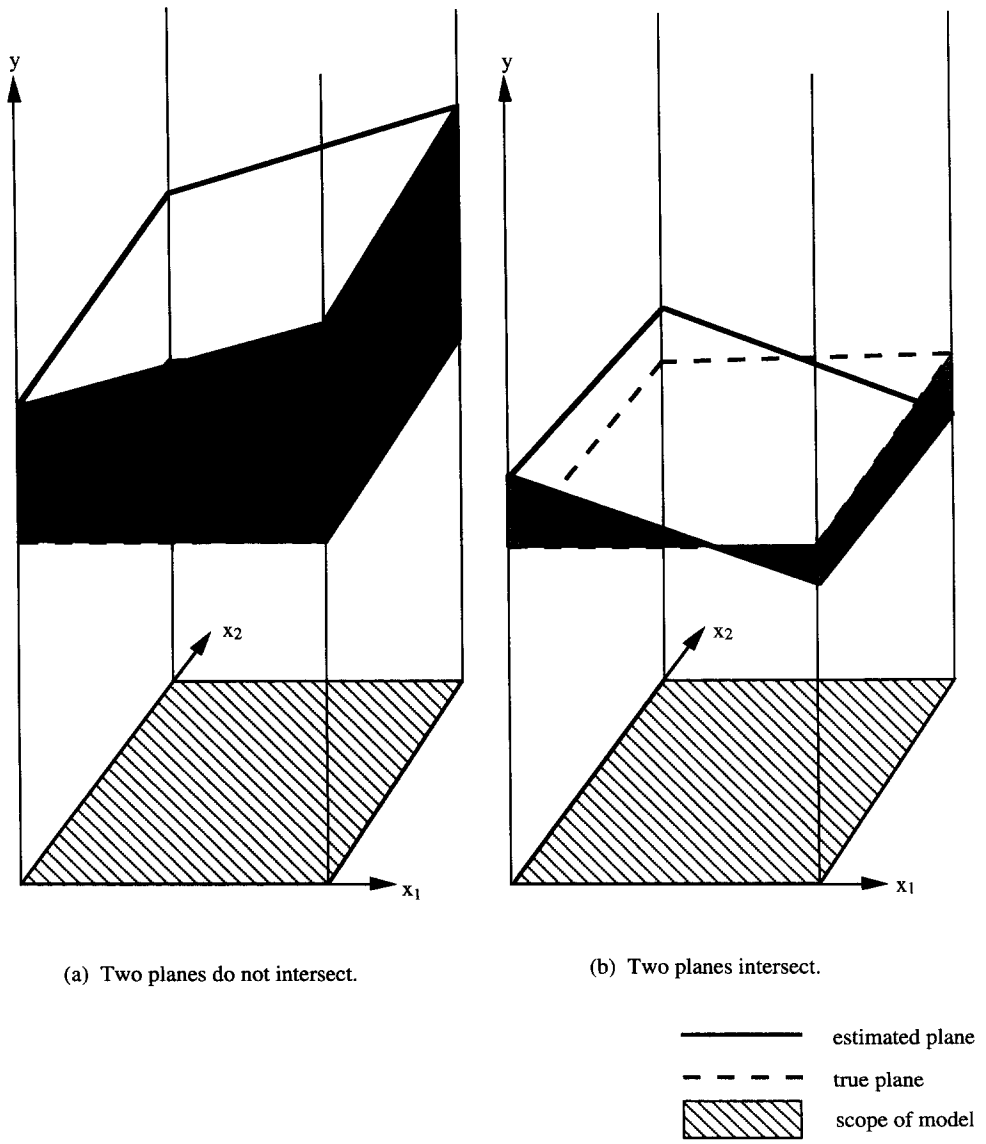
Descriptive capability is concerned with how close the estimated parameters are to the true values of the model parameters, and thus measured by the similarity between the true and the estimated model equations, or hyperplanes. The similarity between two hyperplanes can be evaluated by their angle and spatial closeness (Rogers et al., 1991). Angle closeness is a measure of the degree of parallelism between the two hyperplanes. Spatial closeness is a measure of the distance between the portion of the hyperplanes within the scope of the model.

Since the underlying model given in (3.1) has two independent variables, the true and estimated hyperplanes are planes in three dimensional space. In three dimensional space, the two criteria for similarity were integrated into a single criterion of 'volume' which was formed by the discrepancy between the two planes within the scope of the model. Fig. 2 shows two possible cases; i.e., the two planes may or may not intersect within the scope of the model. (The volume between the two planes is shaded in Fig. 2.) In both cases, an estimated plane yielding a smaller volume is considered closer (or more similar) to the true plane. For example, in Fig. 2, the estimated plane in (b) yields a smaller volume and thus is considered closer to the true plane than that in (a).

3.2.2. Measures of predictive capability: #IN and MSE

After the model was fitted by statistical regression and fuzzy regression, 100 points (i.e., observations) were then randomly generated according to the factor levels of the corresponding treatment. Then, their predictive capabilities were evaluated via the following two measures; #IN and MSE.

#IN: The number of points (out of the randomly generated 100 points) that fell in the estimated intervals from statistical regression and fuzzy regression was counted and compared. For this purpose, a 95% prediction interval for a new observation was used in statistical regression and a fuzzy interval in fuzzy regression. As pointed out earlier, these intervals are very different, and hence, must be adjusted to permit an equitable comparison. To resolve this issue, a value of H, H^*, was found and used for fuzzy regression at which the volume between the upper and lower limit of the H^*-cut of the fuzzy interval was equal to that of the 95% prediction interval within



(a) Two planes do not intersect.

(b) Two planes intersect.

Fig. 2. True and estimated planes in three-dimensional space.

Table 2

	i																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_{i1}	1	5	11	16	20	3	7	10	14	17	2	4	6	8	9	12	13	15	18	19
x_{i2}	11	16	1	20	5	3	17	7	10	14	2	12	18	19	9	4	8	13	6	15

the scope of the model. A detailed procedure for this is given in Appendix A.

MSE (*Mean Squared Error*): MSE is defined as

$$\text{MSE} = \frac{\Sigma_{i=1}^{100}(y_i - \hat{y}_i)^2}{100 - 3}, \tag{3.2}$$

where y_i and $\hat{y}_i$ denoted the i-th observed and estimated value of the dependent variable, respectively, $i = 1, \ldots, 100$. Note that the y_i's in (3.2) are the observations generated after the model was specified and its parameters estimated, not those used to estimate the regression parameters (analogous to a holdout sample). Thus, statistical regression does not necessarily result in smaller MSEs than fuzzy regression, albeit statistical regression obtains parameter estimates by minimizing the sum of the squared errors.

3.3. Experimental procedure

The true parameter values of the model given in (3.1) were assumed to be known. β_0 and β_1 were set at 2 and 1, respectively. The value of β_2 depended on the level of factor MA, and is discussed below. Thus, the regression model from which the data were generated was

$$y_i = 2 + x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \tag{3.3}$$

where ε_i's are normally distributed with mean zero. However, the regression equation which was actually fitted included only the first independent variable (x_{i1}), and intentionally omitted the second independent variable (x_{i2}). Therefore, the fitted regression equation can be represented as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}.$$

In (3.3), x_{i1} and x_{i2} were known constants, given in Table 2. [5]

Now, we discuss how the factor levels were implemented in this experiment.

1) *Number of data points* (N): When N was SMALL, it was set at 5 and 10, respectively, in the experiments NQMHA and NQMHN. [6] When N was LARGE, 20 data points were used in both experiments.

2) *Quality of data set* (Q): The level of dispersion of the data points can be controlled by the variance of the error terms, σ^2. σ^2 was set at 2 when Q was GOOD, and 200 when Q was bad. (Also see the discussion on heteroscedasticity below.)

3) *Aptness of the regression model* (MA): When MA was GOOD, β_2 was set at zero. Thus, the underlying model had only one independent variable. When MA was BAD, β_2 was set at 2. In this case, the underlying model actually had two inde-

[5] The range of x_{i1} for $i = 1$ to 5 was equal to the ranges of x_{i1} for $i = 1$ to 10, and $i = 1$ to 20. Thus, the scope of the model remained constant for all factor levels of N. (Also see the explanation concerning the factor N later in this subsection.) The same held for x_{i2}.

The width of a confidence interval at x_h in statistical regression increases as a function of $|x_h - \bar{x}|$, while that of a fuzzy interval in fuzzy regression increases as a function of $|x_h|$. Therefore, it is more appropriate to compare the (predictive) capabilities of the two methods when $\bar{x}$ coincides with zero. Then, the widths of both intervals are smallest at the same point of x, i.e., at the origin. In the given data set, the values of x_{i1} (and x_{i2}) ranged from 1 to 20. But, in the actual experiment, the observations were shifted horizontally so that the mean of the shifted x_{i1} (and x_{i2}) values became zero.

[6] N was set at 10, instead of 5, in experiment NQMHN because when $N = 5$, even the extreme number of runs has a large chance of occurrence. Therefore, factor NR cannot be implemented meaningfully when $N = 5$. (See Appendix B.)

Table 3
Comparison of performance between statistical and fuzzy regression

Factor	Descriptive		Predictive			
	Measure					
	VOLUME		#IN		MSE	
	Experiment					
	NQMHA	NQMHN	NQMHA	NQMHN	NQMHA	NQMHN
Significant factors [a]						
– main	***N*, *Q*, MA**	***Q*, MA, HS**	***Q*, MA, HS**	***N*, *Q*, MA, HS**	*N*	***Q*, HS**
– interaction	***N* * *Q*, *N* * HS** *N* * AC	***Q* * HS**	***N* * AC, *Q* * MA** ***Q* * HS, MA * HS**	***Q* * MA, *Q* * HS** **MA * HS**	*N* * *Q*, *N* * HS	***Q* * HS**
All factors combined	(–) [c]	S [b]	S	S	(–)	S
N = SMALL	F	S	S	S	F	S
LARGE	S	S	S	S	(–)	S
Q = GOOD	F	(–)	S	S	(–)	S
BAD	(–)	S	S	S	(–)	S
MA = GOOD	S	S	S	S	(–)	S
BAD	(–)	S	S	S	(–)	S
HS = NONE	(–)	(–)	S	S	(–)	S
SEVERE	(–)	S	S	S	(–)	S
AC = NEGATIVE	(–)		S		(–)	
ZERO	(–)		S		(–)	
POSITIVE	(–)		S		(–)	
NR = SMALL		S		S		S
MEDIUM		S		S		S
LARGE		S		S		S
N = *S*, MA = *B*, HS = *S* AC = *N*/*P* or NR = *S*/*L*	F	(–)	S	S	F	(–)
N = *L*, MA = *G*, HS = *N* AC = *Z* or NR = *M*	S	S	S	S	S	S

[a] Factors in bold face were significant at $\alpha = 0.01$, all other factors were significant at $\alpha = 0.05$.
[b] S (F) means that statistical (fuzzy) regression performed better at $\alpha = 0.05$.
[c] (–) means that there was no significant difference in the performances of the two methods at $\alpha = 0.05$.

pendent variables, but the second independent variable (x_{i2}) was omitted from the fitted model. [7]

4) *Heteroscedasticity* (HS): One of the usual assumptions about the error terms is that they are normally distributed with mean zero ($N(0, \sigma_i^2)$). In many real applications, the error variance, σ_i^2, increases with x. Specifically, here it was assumed that the error variance increased exponentially with the first independent variable (x_{i1}). Namely, we let

$$\sigma_i^2 = \sigma^2 \exp(k(x_{i1} - 1)), \tag{3.4}$$

where σ^2 is the variance specified by the level of Q, and k is a constant which determines the rate of increase. The value of k was set at zero when HS was NONE, and 0.242 when HS was SEVERE. Note that $\sigma_i^2 = \sigma^2$ for all x_{i1} if $k = 0$, and $\sigma_i^2 = \sigma^2$ at $x_{i1} = 1$ and $\sigma_i^2 = 100\sigma^2$ at $x_{i1} = 20$ if $k = 0.242$.

5) *Autocorrelation* (AC): To induce correlation between the error terms, a first-order autoregres-

[7] Since the ranges of x_{i1} and x_{i2} were the same, the standardized coefficient of β_1 was approximately half of that of β_2. Thus, x_{i2} (which was omitted from the model) was approximately twice as important as x_{i1}.

sive error model was employed. The error term was expressed as

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i, \tag{3.5}$$

where u_i's are independent normal error terms with mean zero and constant variance specified by the level of Q, and ρ is the autocorrelation coefficient. ρ was set at -0.9, 0, and 0.9 when the level of AC was NEGATIVE, ZERO, and POSITIVE, respectively.

6) *Nonrandomness of error terms* (NR): The error term in the regression model may be positive or negative. A sequence of consecutive positive (or negative) errors is defined as a 'run'. Too few or too many runs in a data set would indicate nonrandom error terms. When N was SMALL (i.e., 10 data points in NQMHN), data sets with 2, 6, and 9 runs were generated, which correspond to the SMALL, MEDIUM, and LARGE level of factor NR, respectively. When N was LARGE (i.e., 20 data points), the data set had 3, 11, and 17 runs for SMALL, MEDIUM, and LARGE NR, respectively. A detailed procedure for the selection of such levels is given in Appendix B.

3.4. Results

The relative performance of statistical versus fuzzy regression was analyzed by observing the difference in each of the performance measures with respect to both methods. Analysis of variance and paired comparison tests were employed to analyze the results at the 5% significance level. Table 3 indicates the factors that had significant effects on the difference in the performances, and also shows which method performed better at various factor levels. Factors N, Q, MA, and HS were significant in most of the cases with respect to the VOLUME (descriptive capability) and #IN (predictive capability) criteria, while only N (in NQMHA) and Q and HS (in NQMHN) were significant with respect to MSE. Factors AC and NR did not show significant effects, in any of the cases, regarding performance differences between statistical and fuzzy regression.

3.4.1. Descriptive capability (VOLUME criterion)

When all the data from the experiments were combined, there was no difference in performance (p-value = 0.64) in NQMHA, and statistical regression performed significantly better in NQMHN. Although several two-way interactions were present, Table 3 also shows which method performed better at each main factor level. [8] In NQMHA, the performance of fuzzy regression was significantly better than that of statistical regression when N was SMALL and Q was GOOD, while statistical regression outperformed fuzzy regression when N was LARGE and MA was GOOD. In NQMHN, statistical regression performed significantly better except when Q was GOOD and HS was NONE. When the combination of the factor levels was unfavorable to statistical regression (i.e., N was SMALL, MA was BAD) and violated the assumptions of statistical regression (i.e., HS was SEVERE, AC was NEGATIVE or POSITIVE, NR was SMALL or LARGE) fuzzy regression performed better than statistical regression in NQMHA, and no difference was observed in NQMHN at the 5% level (p-values ranged from 0.11 to 0.21). However, when the factor levels were favorable to statistical regression (i.e., N was LARGE, MA was GOOD) and satisfied the assumptions of statistical regression (i.e., HS was NONE, AC was ZERO, NR was MEDIUM), statistical regression outperformed fuzzy regression in both experiments.

3.4.2. Predictive capability (#IN and MSE criterion)

Statistical regression performed significantly better than fuzzy regression with respect to the #IN criterion. Fuzzy regression did not perform significantly better than statistical regression in any of the 48 ($= 2^4 * 3$) treatments in both experiments. Among the 48 treatments, statistical regression significantly outperformed fuzzy regression in 12 and 21 treatments in NQMHA and NQMHN, respectively. The mean of the differences (from 20 replications per treatment) of the numbers of observations contained in the prediction and fuzzy intervals was as large as 7.6 in

[8] Checking the interaction structures reveal that the interacting effects did not contradict the main factor level effects shown in Table 3. See Appendix C for the interpretation of selected interactions.

favor of statistical regression in both experiments. The reason for such dominance by statistical regression may be explained by the difference in shapes of the prediction interval and fuzzy interval. Since the width of a fuzzy interval was zero when $x = 0$, and increased with $|x|$ (i.e., the spread of the intercept term was zero), it is highly probable, compared to the case of the prediction interval in statistical regression, that randomly generated points at small $|x|$ values would not fall in the interval. When the assumptions of statistical regression are violated, the performance of statistical regression is degraded (i.e., #IN may be much less that 95), but still better than that of fuzzy regression.

With respect to the MSE criterion, in NQMHA, both methods did not show any significant difference except when N was SMALL, in which case fuzzy regression outperformed statistical regression. However, statistical regression was better at each of the main factor levels in NQMHN. When the combination of the factor levels was unfavorable to statistical regression, fuzzy regression performed better in NQMHA, and no difference was observed in NQMHN. Statistical regression significantly outperformed fuzzy regression in both experiments when the combination of the factor levels was favorable to statistical regression.

3.4.3. *Effects of factor level changes*

The impact of factor level changes on the comparative performances of statistical and fuzzy regression was also investigated. Table 4 shows the results indicating which method's performance improved, in a relative sense, as each main factor level changed. Scheffe's multiple comparison of means (Neter, Wasserman and Kutner, 1985) was employed for this analysis.

In Table 4, the results from all three performance measures (VOLUME, #IN, MSE) are virtually consistent. Statistical regression improved, relative to fuzzy regression, as N became LARGE (i.e., the level of N changed from SMALL to LARGE), Q became BAD, MA became GOOD, and HS became SEVERE. Conversely, fuzzy regression improved, relative to statistical regression, as the factor levels changed in the reverse direction. Factors AC and NR did not have significant effects on the relative performance of the two methods as their levels changed. These results indicate that fuzzy regression can be a useful tool if the data set is insufficient to support

Table 4
Effect of factor level changes

Factor	From → to	Descriptive		Predictive				Overall
		Measure						
		VOLUME		#IN		MSE		
		Experiment						
		NQMHA	NQMHN	NQMHA	NQMHN	NQMHA	NQMHN	
N	SMALL → LARGE	S [a]	(S) [b]	(S)	S	S	(F)	S
Q	BAD → GOOD	F	F	F	F	(F)	F	F
MA	BAD → GOOD	S	S	S	S	(S)	(F)	S
HS	SEVERE → NONE	(F)	F	F	F	(F)	F	F
AC	NEGATIVE → ZERO	(F)		(S)		(F)		–
	POSITIVE → ZERO	(F)		(S)		(F)		–
NR	SMALL → MEDIUM		(S)		(S)		(S)	–
	LARGE → MEDIUM		(S)		(F)		(S)	–

[a] S (F) means that statistical regression (fuzzy regression) improved, in a relative sense, as the factor level changed from the initial level to the last level ($\alpha = 0.05$).

[b] A parenthesis means that the performance change was not significant at $\alpha = 0.05$.

statistical regression analysis and/or the aptness of the model is bad due to vague relationships among variables or poor model specification, where the application of statistical regression is neither appropriate nor effective.

The performance changes resulting from the level changes of factors N and MA conform to our expectation. As the size of the data set enlarges, statistical regression can estimate the regression parameters and prediction intervals with more precision according to statistical theory. On the other hand, more data points do not improve (indeed reduce) the performance of fuzzy regression, in the sense that the estimates from fuzzy regression become fuzzier as the number of data points increases. If the level of MA is BAD (i.e., the regression model is misspecified), the basic assumptions about the error terms in statistical regression are likely to be violated, adversely affecting the validity and performance of statistical regression. As fuzzy regression does not make any assumptions about the error terms, misspecifying the regression model does not directly impact its performance.

In order to understand the results regarding factors Q and HS, it is necessary to reexamine the fuzzy regression algorithm. Fuzzy regression determines the center values and spreads of regression parameters in such a way that *every* observation in the given data set is included in the H-cut of the estimated fuzzy interval. Thus, if there is an outlier (a point lying far beyond the scatter of the remaining points), the parameter estimation of fuzzy regression would be greatly affected. Namely, the resulting parameter estimates would be drastically different from the parameter estimates that would have been obtained without the outlier in the same data set. An outlier would also affect the determination of regression coefficients in statistical regression. However, the contribution of a single point in determining parameter estimates in statistical regression is generally far less than the impact of an outlier in fuzzy regression, and decreases as N increases. Hence, fuzzy regression is more sensitive to outliers or a large variation in the data set than statistical regression. Accordingly, the performance of fuzzy regression is significantly degraded, vis-à-vis statistical regression, when Q is BAD and/or HS is SEVERE because the data set at those factor levels has a large variation, and is likely to have outliers.

4. Discussion

Statistical regression and fuzzy regression have been developed from different perspectives; namely, probability theory and possibility theory, respectively. Because their methods of parameter estimation and therefore estimates are different, a DM has to decide which approach to use in a specific problem situation. This paper conceptually and empirically compared statistical regression with fuzzy regression. This was done to better understand the contrast in mechanisms of the two methods, assess the conditions under which one outperforms the other, and to provide insight regarding the use of a fuzzy methodology in analyzing and interpreting data.

Generally, there are certain circumstances under which the application of fuzzy regression is appropriate, but statistical regression is not – for example, when the observations of the dependent variable are obtained as fuzzy numbers (Tanaka, Uejima and Asai, 1982), or some data are collected by measurements (i.e., crisp), yet other data are estimated qualitatively (i.e., fuzzy) (Johnson and Ayyub, 1993). Such cases often occur when human expert knowledge is the main source of the data or information needed for modeling.

The objectives of regression analysis are to make predictions and describe/explain phenomena. Our experiments showed that in terms of predictive capability, statistical regression is generally superior to fuzzy regression. However, with respect to descriptive capability, performance, and therefore choice of method, depends on various factors associated with the data set, proper specificity of the model, and the nature of the process being investigated. Interestingly and somewhat counter-intuitively, fuzzy regression should not be

used when the data quality is bad (i.e., when there exist outliers or high variability in the data).

Although it is conventional and generally superior to use statistical regression analysis for predicting and describing phenomena from data, there are cases when fuzzy regression analysis is not only appropriate but indeed superior in describing a process. In general, one of the characteristics of the problems suitable for statistical regression is that they are open to experimentation. This, in essence, means that sufficient data must be available for deriving a valid statistical relation among variables. Often, in practice, this is not the case. Another important characteristic of problems amenable to statistical regression is the validity of the assumptions in order for the linear regression model to be considered in practical applications (Savic and Pedrycz, 1991). Fuzzy regression provides a viable, alternative approach for modeling situations which fail to satisfy one or both of the above characteristics.

For example, in quality function deployment which is a process for integrating the 'voice of the customer' into product design, the relationships among the customer attributes (which are usually qualitative) and the engineering characteristics (which are generally quantitative) are typically vague and imprecise in practice because of the general inherent ambiguity in the system. Further, data available for product design is often limited, inaccurate, or vague at best (particularly when developing an entirely new product). There are many such cases in engineering and management problems in practice, which makes fuzzy regression analysis a potentially useful tool. As our experimental results revealed, fuzzy linear regression should be useful and can thus be used as a viable alternative to statistical linear regression in estimating regression parameters when the data set is insufficient to support statistical regression analysis and/or the aptness of the regression model is bad (e.g., due to vague relationships among variables and poor model specification).

As a future research direction, it would be desirable to consider a case that has more than two independent variables to provide more insight. As the number of independent variables increases, this would involve considerations on the general structural issues such as multicollinearity, differences in relative importance of multiple independent variables, etc. It may also be interesting to compare the performance of statistical and fuzzy regression when the underlying model is not linear.

Appendix A. Determination of fuzzy interval

The width of a fuzzy interval $x = x^*$ is

$$2\sum_{j=0}^{2} c_j |x_{ij}^*|,$$

and the width of its H-cut is

$$2|L^{-1}(H)|\sum_{j=0}^{2} c_j |x_{ij}^*|,$$

as can be inferred from (2.2) and (2.3), where x_{i0} is an indicator variable for the intercept term, and is always set equal to 1. (For instance, the distance between $y_{(l)}$ and $y_{(r)}$ in Fig. 1 is the width of the fuzzy interval, and the distance between y' and y'' is the width of its H-cut at $x = x_{i1}$.) The width of a fuzzy interval depends on the H parameter value, and increases with $|x|$. Thus, the width and shape of this interval are different from those of a prediction interval obtained from statistical regression.

In order to make a fair comparison, the volume between the upper and lower limit of the 95% prediction interval was made equal to that of the H-cut of the fuzzy interval via a numerical method. Namely, H^*, an H value which satisfies the following condition, was found:

$$\int_{x_1}\int_{x_2} (y^u - y^l)\, dx_2\, dx_1 = \int_{x_1}\int_{x_2} 2|L^{-1}(H^*)|\sum_{j=0}^{2} c_j |x_{ij}|\, dx_2\, dx_1,$$

where y^u and y^l denote the upper and lower limit of the 95% prediction interval, respectively, from statistical regression. Then, the H^*-cut of

the fuzzy interval which was used for comparison is

$$\left(\sum_{j=0}^{2} \alpha_j x_{ij} - |L^{-1}(H^*)| \sum_{j=0}^{2} c_j |x_{ij}|, \right.$$
$$\left. \sum_{j=0}^{2} \alpha_j x_{ij} + |L^{-1}(H^*)| \sum_{j=0}^{2} c_j |x_{ij}| \right).$$

Appendix B. Selection of the levels of factor NR

In order to see the effect of the nonrandomness, a data set was generated which has the minimum possible, maximum possible, and median number of runs for any (n_1, n_2) pair (by rearranging the order of observations), where n_1 denotes the number of observations above the true value (i.e., the associated error term is positive), and n_2 denotes the number of observations below the true value (i.e., the associated error term is negative). Note that n_1 and n_2 are random variables with the restriction of $n_1 + n_2 = N$. The minimum possible number of runs is equal to 1 if either n_1 or n_2 is zero, and 2 if both n_1 and n_2 are positive. The maximum possible number of runs is equal to

$$\begin{cases} \min(n_1, n_2) * 2 + 1 & \text{if } n_1 \neq n_2, \\ 2n_1 & \text{if } n_1 = n_2. \end{cases}$$

However, for some (n_1, n_2) pairs, even the extreme number of runs has a large chance of occurrence (p-value) when the error terms are actually random. For example, consider a data set of 20 points with $n_1 = 5$ and $n_2 = 15$. The maximum possible number of runs is equal to $5 * 2 + 1 = 11$, and the associated p-value is 0.13, which is quite likely to occur when error terms are random.

To make the number of runs significantly small or large, it is necessary to restrict the set of (n_1, n_2) pairs which would be used in this experiment, i.e., the extreme (or near extreme) number of runs for an eligible (n_1, n_2) pair should have a significantly small p-value. By calculating the p-values of various (n_1, n_2) combinations, the following were selected as the eligible pairs; $(n_1, n_2) = (8, 12), (9, 11), (10, 10), (11, 9), (12, 8)$ for $N = 20$, and $(4, 6), (5, 5), (6, 4)$ for $N = 10$. For example, when $(n_1, n_2) = (8, 12)$, the maximum number of runs is equal to 17, and

$$P(\text{number of runs} = 17) = 0.0013.$$

When $(n_1, n_2) = (4, 6)$, the maximum number of runs is equal to 9, and

$$P(\text{number of runs} = 9) = 0.0238.$$

Both p-values are considered sufficiently small.

Table B.1
Calculation of p-values

N	(n_1, n_2)	p-value	Selected (R_S, R_L)
10	(4, 6) or (6, 4)	$P(R \leq 2) = 0.0095$ $P(R \geq 9) = 0.0238$	(2, 9)
	(5, 5)	$P(R \leq 2) = 0.0079$ $P(R \geq 9) = 0.0397$	
20	(8, 12) or (12, 8)	$P(R \leq 2) = 0.00002$ (< 0.0001) $P(R \leq 3) = 0.0002$ $P(R \geq 17) = 0.0013$	(3, 17)
	(9, 11) or (11, 9)	$P(R \leq 2) = 0.00001$ (< 0.0001) $P(R \leq 3) = 0.0001$ $P(R \geq 17) = 0.0035$	
	(10, 10)	$P(R \leq 2) = 0.00001$ (< 0.0001) $P(R \leq 3) = 0.0001$ $P(R \geq 17) = 0.0045$	

B.1. Selection of significantly small or large number of runs

Let R_S and R_L denote the 'SMALL' and 'LARGE' number of runs (levels 1 and 3 of factor NR) for a given (n_1, n_2) pair, respectively. R_S and R_L are selected by the following criteria:

(a) $P_S^{(10, 10)}$ and $P_L^{(8, 12)}$ are large enough so that Z-values corresponding to $P_S^{(10, 10)}$ and $P_L^{(8, 12)}$ can be found in the standard normal distribution table, i.e., p-values should be at least 0.0001.

(b) $P_S^{(8, 12)}$ and $P_L^{(10, 10)}$ are small enough so that the data set is considered to show a significantly nonrandom pattern.

Here,

$$P_S^{(n_1, n_2)} = P(\text{number of runs} \le R_S),$$

and

$$P_L^{(n_1, n_2)} = P(\text{number of runs} \ge R_L)$$

for an (n_1, n_2) pair. The criterion (a) is required for the large sample approximation, which is explained later in this Appendix.

Note: For (n_1, n_2) and (n_1', n_2') such that $n_1 + n_2 = n_1' + n_2'$ and $|n_1 - n_2| > |n_1' - n_2'|$, $P_S^{(n_1, n_2)} > P_S^{(n_1', n_2')}$ and $P_L^{(n_1, n_2)} < P_L^{(n_1', n_2')}$.

Table B.1 shows the p-values for several cases (Mood, 1940). According to the criteria described above, 2 and 9 were selected as R_S and R_L, respectively, for $N = 10$, and 3 and 17 were selected as R_S and R_L, respectively, for $N = 20$.

B.2. Selection of median number of runs

Let R_M be the median of possible number of runs (level 2 of factor NR) for a given (n_1, n_2) pair, i.e.,

$$P(\text{number of runs} < R_M) \le 0.5$$

and

$$P(\text{number of runs} \le R_M) \ge 0.5.$$

This is equivalent to selecting R_M such that $|P^{(n_1, n_2)} - P^{(n_1, n_2)}|$ is smallest, where

$$P^{(n_1, n_2)} = P(\text{number of runs} \le R_M)$$

and

$$P^{(n_1, n_2)} = P(\text{number of runs} \ge R_M)$$

for a given (n_1, n_2) pair. Such an R_M has the largest p-value. As a result, 6 and 11 were selected as R_M for $N = 10$ and $N = 20$, respectively.

B.3. Large sample approximation

When 100 data points (this data set will be referred to as the test data set) were generated after model fitting to measure predictive capabilities, the test data set was generated in such a way that it follows the same random/nonrandom pattern, by controlling the number of runs, as in the original data set. (Original data set refers to the data set from which the regression parameters were estimated.) Specifically, a test data set was generated to have the number of runs whose p-value is equal to the p-value of the number of runs in the corresponding original data set. The large sample approximation formula (Daniel, 1990, p.66) was employed to determine the required number of runs for the test data set (denoted as R);

$$R = \left(\frac{2m_1m_2}{m_1 + m_2} + 1\right) \pm Z_p\sqrt{\frac{2m_1m_2(2m_1m_2 - m_1 - m_2)}{(m_1 + m_2)^2(m_1 + m_2 - 1)}},$$

where m_1 (m_2) denotes the number of observations above (below) the true value (i.e., the associated error term is positive (negative)) in the test data set. For example, consider an original data set of 20 points with $n_1 = 8$ and $n_2 = 12$. The p-value of R_L is $P(R_L = 17) = 0.0013$. In the test data set, suppose m_1 and m_2 turn out to be 40 and 60, respectively. Then $Z_p = Z_{1-0.0013} = Z_{0.9987} = 3.00$. Then

$$R = \left(\frac{2*40*60}{40 + 60} + 1\right) \pm 3.00\sqrt{\frac{2*40*60(2*40*60 - 40 - 60)}{(40 + 60)^2(40 + 60 - 1)}}$$

$$\cong 63.$$

This means that the probability of having 63 or more runs in a test data set of 100 points with

$m_1 = 40$ and $m_2 = 60$ is 0.0013, which is equal to the p-value of R_L of the original data set with $N = 20$ and $(n_1, n_2) = (8, 12)$. Therefore, the order of observations of the test data set was randomly rearranged until it had 63 runs unless the original sequence of the observations formed 63 runs.

Appendix C. Interpretations of selected interactions

NQMHA (VOLUME)

- When N = SMALL, fuzzy regression was better for any level of Q and HS ($\alpha = 0.05$).
- When N = LARGE, statistical regression was better if Q = BAD or HS = SEVERE ($\alpha = 0.05$).
- The difference in performance of statistical regression and fuzzy regression at Q = GOOD and Q = BAD (also, at HS = NONE and HS = SEVERE) was small when N = SMALL, and was large at N = LARGE. This means that when N = LARGE, statistical regression performed significantly better than fuzzy regression at Q = BAD than at Q = GOOD.

NQMHN (VOLUME)

- When Q = GOOD, fuzzy regression was better if HS = NONE.
- When Q = BAD, statistical regression was better for any level of HS.
- The difference in performance of statistical regression and fuzzy regression at HS = NONE and HS = SEVERE was small when Q = GOOD, and was large at Q = BAD. This means that when Q = BAD, statistical regression performed significantly better than fuzzy regression at HS = SEVERE than at HS = NONE.

NOMHA (#IN)

- Statistical regression was better when Q = BAD (for any level of MA), Q = GOOD (for any level of HS), and HS = SEVERE (for any level of Q and MA).
- When Q = GOOD and MA = BAD, statistical regression improved significantly (compared to fuzzy regression) as Q became BAD or MA became GOOD.

References

Bardossy, A. (1990), "Note on fuzzy regression", *Fuzzy Sets and Systems* 37, 65–75.

Bardossy, A., Bogardi, I., and Duckstein, L. (1990), "Fuzzy regression in hydrology", *Water Resources Research* 26/7, 1497–1508.

Bardossy, A., Bogardi, I., and Kelly W., (1987), "Fuzzy regression for electrical resistivity – Hydraulic conductivity relationships", in: J. Chameau and J.T.P. Yao (eds.), *Proceedings of the North American Fuzzy Information Processing Society Workshop*, 333–352.

Bellman, R.E., and Zadeh, L.A. (1970), "Decision-making in a fuzzy environment", *Management Science* 17, B141–B164.

Daniel, W.W. (1990), *Applied Nonparametric Statistics*, 2nd ed., PWS-Kent, Boston, MA.

Dubois, D., and Prade, H. (1980), *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, New York.

Gharpuray, M., Tanaka, H., Fan, L., and Lai, F. (1986), "Fuzzy linear regression analysis of cellulose hydrolysis", *Chemical Engineering Communications* 41, 299–314.

Heshmaty, B., and Kandal, A. (1985), "Fuzzy linear regression and its applications to forecasting in uncertain environments", *Fuzzy Sets and Systems* 15, 159–191.

Johnson, P., and Ayyub, B. (1993), "Incorporating uncertainty into bridge pier estimations", in: *Proceedings of ISUMA '93*, The Second International Symposium on Uncertainty Modeling and Analysis, University of Maryland, 136–140.

Kim, K., Moskowitz, H., Dhingra, A., and Evans, G. (1995), "Fuzzy multicriteria models and decision support system for quality function deployment", CMME Working Paper Series No. 91-10-1, Purdue University, West Lafayette, IN.

Mood, A. (1940), "The distribution theory of runs", *Annals of Mathematical Statistics* 11, 367–392.

Moskowitz, H., and Kim, K. (1993), "On assessing the H value in fuzzy linear regression", *Fuzzy Sets and Systems* 58, 303–327.

Neter, J., Wasserman, W., and Kutner, M. (1985), *Applied Linear Statistical Models*, 2nd ed., Irwin, Homewood, IL.

Rogers, D., Plante, R., Wong, R., and Evans, J. (1991), "Aggregation and disaggregation techniques and methodology in optimization", *Operations Research* 39/4, 553–582.

Savic, D., and Pedrycz, W. (1991), "Evaluation of fuzzy linear regression models", *Fuzzy Sets and Systems* 39, 51–63.

Tanaka, H. (1987), "Fuzzy data analysis of possibilistic linear models", *Fuzzy Sets and Systems* 24, 363–375.

Tanaka, H., and Watada, J. (1988), "Possibilistic linear systems and their application to the linear regression model", *Fuzzy Sets and Systems* 27, 275–289.

Tanaka, H., Uejima, S., and Asai, K. (1982), "Linear regression analysis with fuzzy model", *IEEE Transactions on SMC* 12, 903–907.

Zadeh, L.A. (1965), "Fuzzy sets", *Information and Control* 8, 338–353.

Zadeh, L.A. (1975), "The concept of a linguistic variable and its application to approximate reasoning – I", *Information Science* 8, 199–249.