Fitting Functions to Data with Error Bounds: Fuzzy regression with ERRGO

T. Solymosi, J. Dombi
Research Group on the theory of Automata, Szeged

1992

Abstract

Curve fitting problems on the plane, where errors can occur in both variables, tolerance regions can be specified around the data points and only such functions are interest whose curves cross each other of these regions, can be formulated as fuzzy regressions. The usual parameter estimation methods do not guarantee fits within prescribed pointwise error bounds. A univariate fuzzy regression problem is formulated where functions with crisp parameter values are fitted to fuzzy data. The minimum possibility of that the data and the related fitted fuzzy numbers are equal is used to measure the goodness of fit, and is to be maximized over the parameter region. Conditions are discussed under which there is a best fit and an algorithm is proposed to approximate it in the linear case. Due to the reduction to a sequence of basic problems the same method can be used for linear and linearizable functions.

Keywords: fuzzy regression, curve fitting, ERRGO

1 Introduction

Function fitting problems arise frequently, e. g. regression is one of the most often used statistical procedure, but none of the well-known estimation methods guarantee fits within prescribed pointwise error bounds. In the method of least squares a global measure of goodness of fit is minimized, but we can get large deviations for some points although the average error is optimal.

There are scientific measurement situations where we have only few data and we can not make realistic distributional assumptions (so statistical methods are out of consideration), further, errors can occur in both quantities (inaccuracy in setting the explanatory factor is comparable to measurement error in the response variable), moreover, we might want to specify tolerance regions for the data points and we look for functions which are within these errors for each point.

The above mentioned situations can be formulated as fuzzy regression problems, where we have fuzzy data and we assume crisp functional relationship (the parameters are crisp) between the X- and Y-data. This type of fuzzification of the regression problem (fuzzy data, crisp parameters) was considered by BANDEMER (1988)[1], but his estimation method is rather theoretical promising little success in practice. A different model (crisp X-data, fuzzy Y-data, fuzzy parameters) was investigated by TANAKA and WATADA (1988)[3] in the linear case. They estimate the fuzzy parameters of a linear function by solving linear programming problems, so their method is numerically performable.

In this paper we formulate a linear fuzzy regression problem (only for the univariate case), where we have fuzzy X-, and Y-data and we fit a crisp function. The membership functions can be symmetric or asymmetric, of bounded or of unbounded support, different for X- and for Y-data. For a given tolerance level l-H, the H-level intervals define tolerance regions for the points and we look for functions which cross each of these regions. Naturally, we are interested in the smallest tolerance level for which we still get such solution. This is the same as to say that we maximize the level of possibility of the fuzzy fitted values and the fuzzy Y-data. Systems of linear inequalities determine the feasible parameter values for a fixed tolerance level, so we can give a tolerance region for the data follow a probability distribution.

The program ERRGO is a realization of the algorithm to solve the described fuzzy regression problem. It is able to fit several nonlinear functions of two parameters (e.g. exponential, logarithm, power, linear in the reciprocal of X and/or Y). It is fully integrated to Lotus 1-2-3 so the easy communication is guaranteed.

2 Preliminaries

A fuzzy number X is given by its membership function

$$m_X : \mathrm{Re} = (-\infty, +\infty) \longrightarrow [0, 1]$$

which assigns to each real number the level of that it belongs to the fuzzy number. The set of real numbers which belong to X with at least h level of membership is called the h-level cut of X, with the only exception of 0-level cut, which is the closure of the set of the numbers with positive membership values.

We demand the following properties on fuzzy number X:

(i) $[X]_h$ is a closed, bounded interval for every $0 < h \leq 1$,

(ii) $[X]_1$ is a singleton (called the apex of X),

(iii) functions $lx(h)$, $rx(h)$ defined by $[X]_h = [lx(h),\ rx(h)]$ are continuous on $[0, 1]$

Remark 1 *If $[X]_0$ is unbounded then $lx(0) = \lim_{h\to 0+} lx(h) = -\infty$ and $rx(0) = \lim_{h\to 0+} rx(h) = +\infty$ are possible. For every h $lx(h) \leq x_0 \leq rx(h)$ holds, where x_0 denotes the apex of X. Function $lx(h)$ is increasing $rx(h)$ is decreasing in h.*

An equivalent definition with explicit properties for the membership function can be given as follows: $m_X(.)$ is

(i) upper-semicontinuous ($[X]_h$ is closed for all h); quasiconcave (for all y, z and $x \in (y, z) : m_X(x) \geq \min\{m_X(y), m_X(z)\}$); $\lim_{x \to \pm\infty} m_X(x) = 0$;

(ii) unimodal (there exist a unique $x_0 : m_X(x_0) - \sup_{x \in \mathrm{Re}} m_X(x)$); normal (there exists $x_0 : m_X(x_0) = 1$);

(iii) strictly monotone in $[X]_0$, increasing on the left and decreasing on the right side of the apex.

The operations on real numbers are extended to fuzzy numbers by the extension principle introduced by ZADEH[4]. Real number y belongs to the extended value $f(X)$ of univariate function $f(.)$ for fuzzy number X with level of membership

$$m_{f(X)}(y) = \sup\{m_X(x) : f(x) = y\}.$$

If $f(.)$ is continuous then its extended value for fuzzy number X is also a fuzzy number according to the above definition. It comes easily from the fact that an h-level cut of $f(X)$ is the image of the h-level cut of X by function f. More precisely,

$$[f(X)]_h = f([X]_h) = \left[\min_{x \in [lx(h), rx(h)]} f(x), \max_{x \in [lx(h), rx(h)]} f(x)\right].$$

If $[X]_0$ is unbounded then min, max has to be replaced with inf, sup, respectively.

Since the linear function $f(x) = ax + b$ is monotone, so the interval endpoints of $[ax + b]_h$ are the images of the interval endpoints of $[X]_h$. Explicity, if

$$[X]_h = [lx(h), rx(h)] = [mx(h) - dx(h), mx(h) + dx(h)]$$

where

$$mx(h) = (lx(h) + rx(h))/2$$

denotes the midpoint and

$$dx(h) = (rx(h) - lx(h))/2$$

the half length, then

$$[f(X)]_h = [amx(h) + b - |a|\,dx(h)\ , amx(h) + b + |a|\,dx(h)].$$

In this paper the concept of possibility is used to measure the equality of two fuzzy numbers. The possibility of that fuzzy numbers X and Y are equal is defined by

$$Pos(X, Y) = \sup\{h \in [0, 1] : [X]_h \cap [Y]_h = \emptyset\}.$$

Remark 2 *For an increasing sequence of h values the related intersections from a decreasing (for set inclusion) sequence of nonempty closed intervals, so the supremum can be replaced with maximum in the definition.*

A alternative expression can be given by using the membership functions. We get the well-known formula

$$Pos\,(X, Y) = \sup\{\min\{m_X(z), m_Y(z)\} : z \in \mathrm{Re}\}.$$

Since the membership function of a fuzzy number is strictly quasiconcave (quasi-concave and for all $y, z, x \in (y, z) : m(y) < m(z)$ implies $m(y) < m(x)$) on its support and unimodal, so it is easy to see that $\min\{m_X(.), m_Y(.)\}$ is strictly quasiconcave on $[X]_0 \cap [Y]_0$, unimodal, and its maximum is attained in the closed interval formed by the apexes.

In a function fitting problem several pairs of fuzzy numbers have to be compared and these measures have to be combined into overall criterion. For function $f(.)$ the ggodness of fit to fuzzy data $(X_1, Y_1), ..., (X_{N,}Y_N)$ is defined as the minimum, for all pairs, of the possibilities of that the data Y_j and the related fitted $f(X_j)$ fuzzy numbers are equal, i.e. $\min_{1\le j\le N} Pos\,(Y_j, f(X_j))$.

3 The function fitting problem

We give $N\,(\ge 2)$ pairs of related fuzzy numbers $(X_1, Y_1), ..., (X_N, Y_N)$ as data, and a family of univariate functions of two parameters $\{f(x; a, b) : (a, b) \in M\}$. Let any member of the family be defined on the same convex domain D Re, continuous and monotone (increasing or decreasing) in its variable x on D. Let the set of parameters M Re$\,x\,$Re beconvex. It follows from the continuity of $f(.; a, b)$ that for a fuzzy number X with $[X]_0$ D the fitted value $f(X; a, b)$ is also a fuzzy number. The problem is to find crisp parameters $(a, b) \in M$ such that the related function $f(x; a, b)$ maximizes the goodness of fit to the fuzzy data within the family. Formally

$$\max_{(a,b)\in M}\ \min_{1\le j\le N} Pos\,(Y_j, f(X_j; a, b)).$$

Since the objective function assumes values in $[0, 1]$, so the supremum of its range is also from the unit interval and there are functions yielding arbitrarily good approximations of this theoretical limit.

Introducing the upper sets

$$M(h) = \{(a, b) \in M : Pos\,(Y_j, f(X_j; a, b)) \ge h \qquad j = 1, ..., N\}$$

the problem can be written as

$$\text{Find } \sup\{h \in [0, 1] : M(h) = \emptyset\}.$$

Naturally, $M(h)$ is decreasing (for set inclusion) in h and $M(0) = M$.

If for any h we can decide whether $M(h)$ is empty or not, then we can approximate the supremum by successively halving the closed intervals containing it. If $M(1) = \emptyset$ then the absolute maximum is reached, otherwise, since $M(0) = \emptyset$ so in each step we have an interval $[lh_k, rh_k]$ with $M(lh_k) = \emptyset$, $M(rh_k) = \emptyset$. The number of iterations needed for the prescribed accuracy $epsh > 0$ is $int(-\ln(epsh)/\ln(2)) + 1$.

A more interesting question is how to solve the basic problem, that is to indicate if $M(h)$ is empty, or to find a point of $M(h)$, or to give the whole set. Being able to solve the problem of providing the whole $M(h)$ a tolerance region can be determined for the parameters at each level. In this way the fuzziness in the data can be translated into the parameter space, similary to statistics, where probability distributions are used to represent imprecision, uncertainty in data and different level confidence regions can be specified for the estimates. In both cases the assumed parameters and their estimates are exact numbers, but errors in data carries over into the solution.

The function fitting problem can be interpreted geometrically as follows. The h-level cuts $[X_j]_h, [Y_j]_h \quad j = 1, ..., N$ of the fuzzy data are N pairs of compact intervals. The continuous function $f(x; a, b)$ translates $[X_j]_h$ into its image interval $[f(X_j; a, b)]_h$. To get $Pos(Y_j, f(X_j; a, b)) \geq h$ the fitted interval $[f(X_j; a, b)]_h$ has to intersect the data interval $[Y_j]_h$. This is the same as requiring that the graph of the function has to cross the compact rectangle $[X_j]_h \, x \, [Y_j]_h$. Thus, to guarantee an at least h-level fit simultaneously for all pairs the graph of the function has to cross each of these rectangles. By increasing/decreasing h the rectangles get smaller/larger. The problem is to find the highest value and a function whose graph crosses each of the corresponding smallest rectangles.

Let us investigate the conditions under which we can claim more about the above optimization problem and the upper sets. Let $f(x; a, b)$ be a bevariate continuous and quasimonotone (quasiconvex and quasiconcave) function of parameters (a, b) on M for any fixed x. Let the parameters of the increasing (decreasing) constant (strictly increasing) strictly decreasing functions form the convex set $M_+ \ /M_- \ M_0 \ SM_+ \ SM_-$.

It follows the continuity of $f(.; a, b)$ that for a fuzzy number X with $[X]_0$ D the fitted value $f(X; a, b)$ is also a fuzzy number, thus its membership function $m_{f(X;a,b)}(.)$ is upper-semicontinuous and strictly quasiconcave on its support. On the other hand, it can also be shown that $m_{f(X;a,b)}(z)$ is upper-semicontinuous in $(a, b) \in M$ for any fixed z. Unfortunately, for fixed z the quasiconcavity in (a, b) for $m_{f(X;a,b)}(z)$ can be claimed only on M_+ (M_-) but not on M, and strict quasiconcavity only for points of the intersection of the supports, only on SM_+ (SM_-) and only if $f(x; ., .)$ is strictly quasimonotone for any fixed x.

Now let us consider $Pos(Y, f(X; a, b))$ as a function of (a, b). The proof of its upper-semicontinuity in $(a, b) \in M$ is basically the same as the one for the fitted membership function. Analogously, if increasing and decreasing functions are not mixed, i.e. on M_+ (or on M_-), then the quasiconcavity can be shown also in a similar way. Finally, the strict quasiconcavity on its support restricted to SM_+ (SM_-) also holds, provided that $f(x;.,.)$ is strictly quasimonotone for any fixed x.

To conclude the discussion on the goodness of fit as a function of the parameters we recall (see e.g. MARTOS (1975)[2]) that the minimum of a finite number of upper-semicontinuous/quasiconcave/strictly quasiconcave functions is also upper-semicontinuous/quasiconcave/strictly quasiconcave. It follows that $\min_{1\leq j\leq N} Pos\left(Y_j, f\left(X_j; a, b\right)\right)$ is upper-semicontinuous in $(a, b) \in M$, quasiconcave on M_+ (M_-), and strictly quasiconcave on its support restricted to SM_+ (SM_-), provided that $f(x;.,.)$ is strictly quasimonotone for fixed x.

If the maximization can be restricted to a compact set of parameters then the upper-semicontinuous objective function attains its supremum, i.e. there is at least one function yielding the best fit possible. This is not necessarily the case, since to data $(X.Y_1), (X, Y_2)$ (the apexes are on a vertical line) no linear function with finite slope can be fitted exactly, but the supremum 1 can be approximated arbitrarily well.

For any h value $M(h)$ is a closed set because of the upper-semicontinuity of the objective function, further, $M_+(h) = M(h) \cap M_+$ is convex and closed due to the quasiconcavity on M_+.

4 The linear case

If the functions to be fitted are the linear ones $f(x; a, b) = ax + b$ then $D = \mathrm{Re}$, $M = \mathrm{Re}\, x\, \mathrm{Re}$, $M_{+,-,0} = \{a \geq, \leq, = 0\}$ and clearly, $f(.; a, b)$ is strictly monotone or constant for any $(a, b) \in M$, and $f(x;.,.)$ is strictly quasimonotone for any $x \in D$, in addition to the continuity of both functions.

The function fitting problem can be stated as follows. The crisp parameters of that linear function are to be found, for which the possibility simultaneously guarantable for every pair of data and fitted values is the highest, i.e. $\max_{(a,b,h)\in Mx[0,1]} h$ subject to $Pos\left(Y_j, aX_j + b\right) \geq h \quad j = 1, ..., N$

If the supremum of levels for which the level cuts have nonempty intersections is not smaller than a given level, then the intersection of the cuts for the given level in nonempty, and vice versa. Thus we can equivalently write

$$\max_{(a,b,h)\in Mx[0,1]} h \text{ subject to } Pos\left[Y_j\right]_h \cap \left[aX_j + b\right]_h = \emptyset \quad j = 1, ..., N$$

The intersection of two closed ontervals is nonempty if and only if the left endpoints of the intervals are not greater than the right ones.

Therefore, with the above formula for $\left[aX_j + b\right]_h$ and with $[Y]_j = \left[ly_j(h), ry_j(h)\right]$ we get

$$\max_{(a,b,h)\in Mx[0,1]} h \quad \text{s.t.}$$

$$amx_j(h) + b - |a|\, dx_j(h) \leq ry_j(h)$$

$$amx_j(h) + b - |a|\, dx_j(h) \geq ly_j(h) \quad j = 1, ..., N$$

We can not solve this nonlinear optimization problem in its generality. Even if only triangular fuzzy numbers are used and so all the $mx(h), dx(h)$ functions are linear in h, but the appearance of the absolute value of the slope parameter in the constraints causes serious difficulties. To avoid the absolute values (in general the nonconvexity of the feasible parameter set) we can split the problem into two subproblems by restricting the region to M_+ and to M_-. On M_+, for example, we get the more attractive setting

$$\max_{(a,b,h)\in Mx[0,1]} h \quad \text{s.t.}$$

$$alx_j(h) + b \le ry_j(h)$$

$$arx_j(h) + b \ge ly_j(h) \quad j = 1, ..., N$$

Let us suppose that the fuzzy numbers are enough nice to make the problem solvable with some nonlinear method, but still we can not provide the whole feasible parameter set for a given h level, moreover, the nonlinear method can be quite difficult to perform. Therefore let us be satisfied with an approximate solution and focus on the basic problem to give $M(h)$ for fixed h level.

Since $M(h) = M_+(h)\ M_-(h)$ so it is enough to solve the subproblems. For fixed h, both sets are determined by systems of linear nonstrict inequalities, i.e. they are polytopes. To find the vertices of $M_+(h) = M(h) \cap \{a \ge 0\}$, for example, we need its lower and upper boundaries on the (a, b) parameter plane. The lower boundary is the graph of

$$b = \max_{1\le j\le N} (-rx_j)\, a + ly_j,$$

the upper one is given by

$$b = \min_{1\le j\le N} (-lx_j)\, a + ry_j.$$

They originate from $(0, \max_{1\le j\le N}\ ly_j), (0, \min_{1\le j\le N}\ ry_j)$ and the linear segments from convex, concave shapes, respectively. To get the at most N vertices of the lower boundary of $M_+(h)$ let us take the functions $b = (-rx_j)\, a + ly_j$ into account primarily according to the decreasing order of ly_j's and secondarily according to the increasing order of $(-ry_j)$'s. If $ly_1 \ge ly_2$ and $(-ry_1) \ge (-ry_2)$ then the second inequality is redundant, so it can be dropped, otherwise, the two functions intersect at

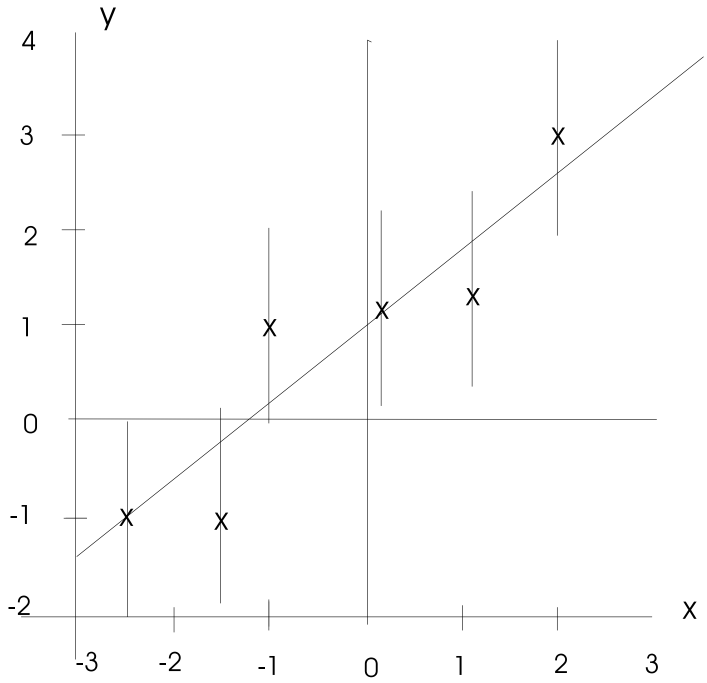
$$\begin{aligned} a &= (ly_1 \ge ly_2) / (rx_1 - rx_2) > 0, \\ b &= (rx_1 ly_2 - rx_2 ly_1) / (rx_1 - rx_2). \end{aligned}$$

This discussion shows that the determination of the boundaries is rather sorting and organization than calculation, since at most $N(N-1)/2$ intersections have to be computed. The same remark can be done for the combination of the lower and upper boundaries. We can again utilize that on the closed halfplane $\{a \ge 0\}$ the vertices can be ordered according to their distances from the $a = 0$ axis.

Types of errors

X : no error   Y : automatic

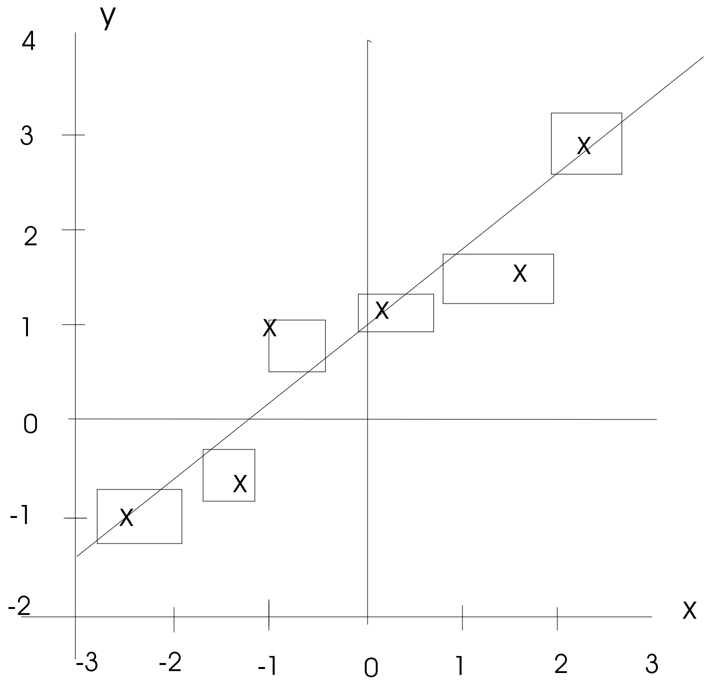
Figure 1.



Tolerance rectangles

Fitted line crosses every rectangle

Figure 2.



5 The case of linearizable functions

Repeating the argument applied in the linear case the upper set

$$M(h) = \left\{(a,b) \in M : [Y_j]_h \cap [f(X_j;a,b)]_h = \emptyset \quad j = 1,...,N\right\}$$

can be written as

$$\begin{aligned} M(h) &= \{(a,b) \in M : lfx_j(h;a,b) \leq ry_j(h) \quad j = 1,...,N\} \\ M(h) &= \{(a,b) \in M : rfx_j(h;a,b) \geq ly_j(h) \quad j = 1,...,N\} \end{aligned}$$

where $lfx_j(h;a,b) / rfx_j(h;a,b)$ is the left/right endpoint of the h-level cut of the fitted value. On M_+ it can be replaced with $f(lx_j(h);a,b) / f(rx_j(h);a,b)$ on M_- with $f(rx_j(h);a,b) / f(lx_j(h);a,b)$ on M_0 with either one.

If there are strictly monotone functions $u(.), v(.)$ such that $y = f(x;a,b)$ can be linearized to $v = c(a,)\,u + d(a,b)$ with transformations $u = u(x), v = v(y)$, further, if the new parameters (c,d) are invertible functions of the original ones and (c,d) maps M onto $\mathrm{Re}\, x\, \mathrm{Re}$ (the parameter set of the linear functions), then the nonlinear basic problem can be solved by translating it to an equivalent linear basic problem.

For fixed h let us transform the h-level cuts of the data

$$[X_j]_h = [lx_j, rx_j], \ \ [Y_j]_h = [ly_j, ry_j]$$

by functions $u(.), v(.)$, respectively. We obtain $[lu_j, ru_j]$, $[lv_j, rv_j]$. Consider the linear basic problem. Find

$$\begin{aligned} \{(c,d) &\in \mathrm{Re}\, x\, \mathrm{Re} : clu_j + d \leq rv_j \quad j = 1, ..., N\} \\ \{(c,d) &\in \mathrm{Re}\, x\, \mathrm{Re} : cru_j + d \geq lv_j \quad j = 1, ..., N\} \end{aligned}$$

To see that this is equivalent to the nonlinear basic problem, let us observe that if $(a,b) \in M(h)$ then on M_+ for a strictly increasing $v(.)$ $f(lx_j; a, b) \leq ry_j$ implies

$$v(f(lx_j; a, b)) \leq v(ry_j) = rv_j.$$

But

$$v(f(lx_j; a, b)) = c(a,b)\, u(lx_j) + d(a,b)$$

thus for a strictly increasing $u(.)$ (for which $lu_j = u(lx_j)$) we get

$$v(f(lx_j; a, b)) \leq rv_j \text{ implies } clu_j + d \leq rv_j.$$

The above implications can be reversed by using the inverses of $u(.), v(.)$ (also strictly monotone) and $a = a(c,d)$, $b = b(c,d)$.

As an example let us consider the family of power functions $f(x; a, b) = ax^b$, $D = (0, \infty)$, $M = (0, \infty)\, x\, \mathrm{Re}$. With strictly increasing transformations $u = \ln(x)$ and $v = \ln(y)$ (if $a > 0$ and $x > 0$ then $y = ax^b > 0$ for any b) we obtain the linearized form $\ln(y) = b \ln(x) + \ln(a)$. For the parameters we have $(c,d) = (b, \ln(a))$ which translates M onto $\mathrm{Re}\, x\, \mathrm{Re}$ in an invertible way $((a,b) = (\exp(d), c))$. If the h-level basic problem is transformable, i.e. $[X_j]_h (0, \infty)$ (domain of $u(.)$) and $[Y_j]_h (0, \infty)$ (domain of v(.)) for all j, then take the rectangles

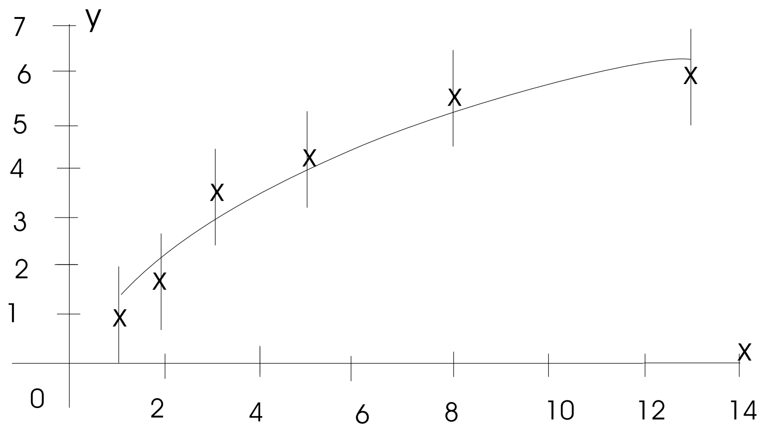
$$[\ln(lx_j), \ln(rx_j)] \ \ x \ \ [\ln(ly_j), \ln(ry_j)] \quad j = 1, ..., N$$

and fit a linear function to them. If $v = cu + d$ is a solution, then $(\exp(d), c)$ is a solution of the power problem, since

$$y = \exp(v) = \exp(cu + d) = \exp(d)(\exp(u))^c = ax^b.$$

Power

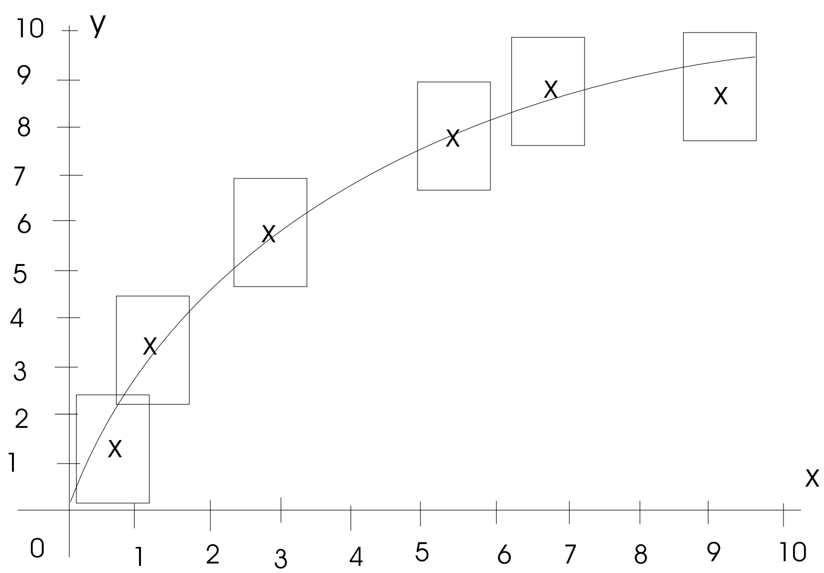
Figure 3.



Nonlinear problem

Fitting power function

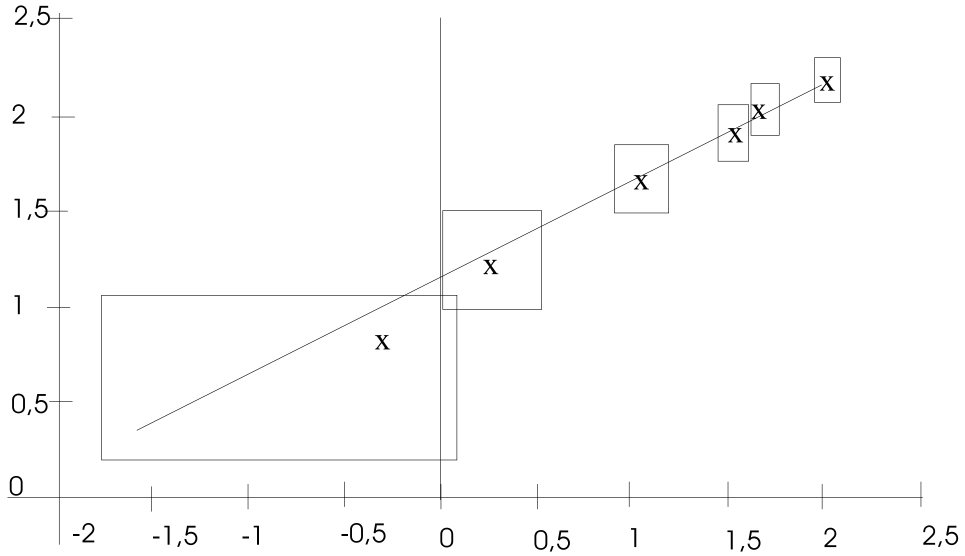
Figure 4.



Transformed problem

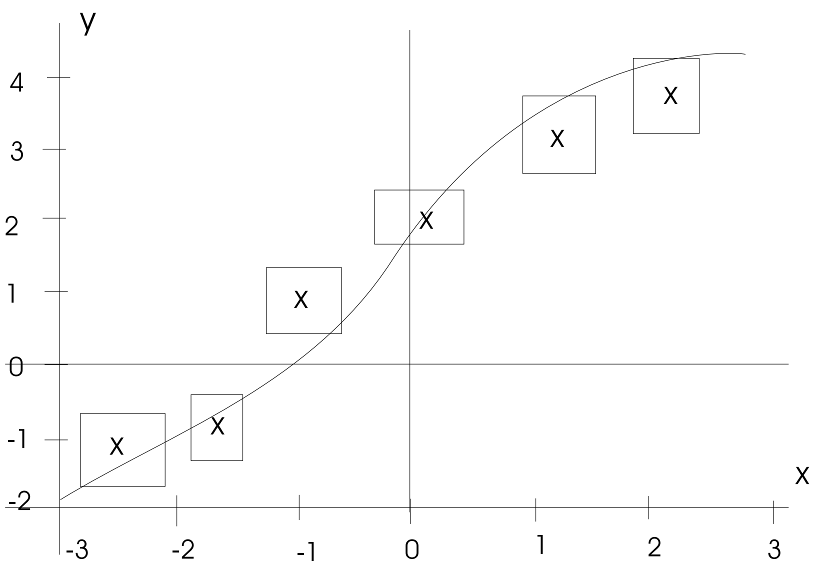
Fitting linear function

Figure 5.



Logistic

Figure 6.



References

[1] H. Bandemer. Evaluating explicit functional relationships from fuzzy observations. *Fuzzy Sets and Systems*, 16:41–52, 1985.

[2] B. Martos. *Nonlinear Programming: Theory and Methods.* Akademiai Kiado, Budapest, 1975.

[3] Watada J. Tanaka, H. Possibilistic linear systems and their application to linear regression model. *Fuzzy Sets and Systems*, 27:275–289, 1988.

[4] L. A. Zadeh. The concept of linguistic variable and its application to approximate reasoning. *Information Science*, 8:199–249, 1975.