

A short primer on fuzzy regression

Janos G. Hajagos
Department of Ecology and Evolution
Stony Brook University
New York, U.S.A.

May 26, 2005

Traditional statistical tools for performing regression cannot handle the large uncertainties present in a fuzzy number dataset. Therefore we will have to rely on techniques developed for interval analysis (Moore 1966; Moore 1979). A useful property of algorithms that handle fuzzy numbers is that they can be rewritten in terms of intervals allowing the powerful mathematical tools of interval analysis to be applied. Every fuzzy number problem can be approached as a level-wise interval problem. The only data requirement is that the fuzzy numbers have the same number of nested intervals. See: Kaufmann & Gupta (1991) for a formal definition of fuzzy numbers.

This primer will focus on estimating bounds on coefficients for a linear model when the dependent variable is composed of measurements which are fuzzy numbers and the independent variable is measured exactly. The feasibility of computing bounds on the coefficients when the independent variable is also inexact will be explored.

1 Least squares approaches

A widely used regression technique for fitting a model to data is least squares regression. In its simplest bivariate form data is fit to the equation for the linear equation $y = mx + b$, where m is the slope and b is the intercept. If the measured dependent variable is an interval then fuzzy least squares regression can be applied (Salia & Ferson 1998).

All least squares linear regression problems can be expressed in terms of matrix computation:

$$\mathbf{y} = \mathbf{XB} + \epsilon,$$

where \mathbf{y} is a $n \times 1$ vector of measured responses, \mathbf{B} is a $m \times 1$ vector of coefficients, and \mathbf{X} is a $n \times m$ matrix of independent measurements or design matrix, and ϵ vector of $n \times 1$ of errors. The least squares solution to the linear model is:

$$\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

If the measurements are exact, that is, they are not fuzzy numbers then computing \mathbf{B} is straightforward. It only requires that a scalar matrix be multiplied by an interval vector. Multiplying an interval vector by a matrix can be defined by decomposing the matrix calculation into binary interval operations (See Neumaier (1990) for definitions of interval matrix operators).

1.1 Example

A hypothetical example will be developed to demonstrate interval least squares regression. Dataset 1 has uncertain dependent variable y and exact independent variable x :

x	y
1	[20, 21]
3	[30, 31]
5	[25, 26]
6	[33, 34]

The first step in computing $\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ is finding the inverse:

$$(\mathbf{X}^T \mathbf{X})^{-1} = \left(\left(\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 6 \end{pmatrix} \right) \right)^{-1} = \begin{pmatrix} 4 & 16 \\ 16 & 84 \end{pmatrix}^{-1} = \begin{pmatrix} 1.05 & -0.2 \\ -0.2 & 0.05 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 1.05 & -0.2 \\ -0.2 & 0.05 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 \end{pmatrix} \begin{pmatrix} [20, 21] \\ [30, 31] \\ [25, 26] \\ [33, 34] \end{pmatrix} = \begin{pmatrix} [19.84, 21.56] \\ [1.49, 1.91] \end{pmatrix}$$

The intercept and slope of the best fit line is, respectively, [19.84, 21.56] and [1.49, 1.91]. The solution set is the interval hull (Marino & Palumbo 2003). The interval hull is the smallest interval vector which encloses the solution set. A solution set is not always rectangular but can sometimes be shaped like a star as in the solution to a pair of linear equations with uncertain coefficients (Schäfer 2004).

2 Parameter bounding

Parameter bounding offers a different approach to estimation than least squares regression. Rather than finding the best fit of a function through a cloud of data points, parameter bounding gives the set of parameter values that intersects each datum. To make this more clear, with regression on exact data for each k parameters in the model we obtain k numerical values but with parameters bounding on interval data we obtain a set in k dimensional space. If a vector is taken out of the k dimensional set and, then, used to parameterize the model each datum in the cloud will be intersected. For many datasets the parameter space will be the null set \emptyset or empty set $\{\}$. If the variance is high and the interval bounds on y are narrow there will not be enough flexibility in a monotonic function to allow it to intersect each datum.

Start with a simple statistical linear model:

$$y_i = \alpha x_i + \beta + \epsilon_i,$$

where x_i is the independent variable known precisely, y_i is the dependent variable, α is the slope of the line, and β is intercept. The ϵ_i , the error term, says that there will be error in the process that generated y_i . In the statistical model each individual ϵ_i is independent and expected to be normally distributed. Because ϵ_i is normally distributed there are no bounds on the value of the ϵ ; of course, in most cases really large error terms are probabilistically unlikely. For a model fit by parameter bounding we assume that ϵ_i is bounded. In addition, we need not make any assumptions on how ϵ_i is distributed. When a parameter bounding algorithm is applied to data with a linear model, we get a 2-dimensional set along the α -axis and β -axis on the real plane. If a pair of values is selected from this set, (α_j, β_j) , and used as coefficients for the linear equation, then the following inequalities will hold:

$$\begin{aligned} y_i - \epsilon_i &\geq \alpha_j x_i + \beta_j \\ y_i + \epsilon_i &\leq \alpha_j x_i + \beta_j \end{aligned} .$$

Or, said more simply, each line we draw having a slope α_j and an intercept β_j will cross the horizontal line at x_i having a midpoint y_i with a width $2\epsilon_i$.

There exists several different methods for parameter bounding. If the model is linear there are exact methods for defining the set of feasible solutions using polytopes to bound the set (Norton 2002). The algorithm for set inversion via interval analysis (Moore 1992; Jaulin 1993) can approximate the parameter bounding set for linear and non-linear models with a

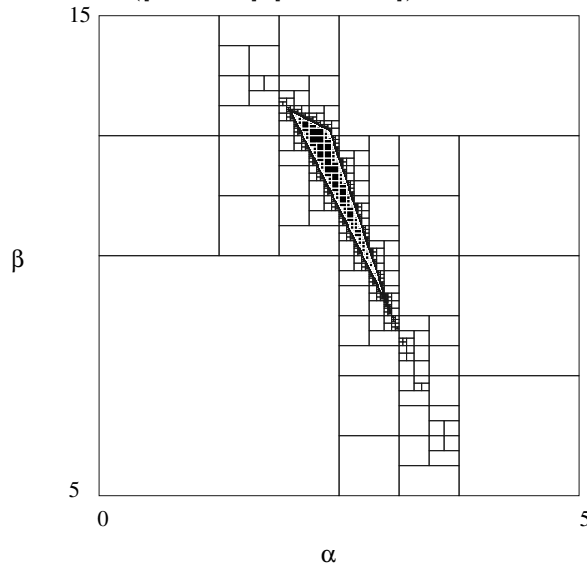
subpaving. A subpaving is a set of interval vectors whose union encloses the solution set. The outer approximation to the feasible set can be made as fine as needed (Jaulin *et al.* 2001).

The answer for performing a parameter bounding on Dataset 1 with a simple linear model is the null set. This proves that no single linear line can be drawn to intersect every datum. Dataset 2 is composed of 4 intervals measured exactly in time.

x	y
1	[10, 15]
3	[16, 20]
4	[21, 25]
6	[22, 27]

The feasible set for α and β is shown in Figure 1.

Figure 1: Poster feasible set for Dataset 2 given a simple linear model. The interval hull solution is $([2.0, 2.99], [9.0, 12.99])$



3 The dependent variable is not exact

The computations when the dependent variable is a fuzzy number is computationally more complex. For least squares regression involves calculating

$(\mathbf{X}^T \mathbf{X})^{-1}$. Calculating an inverse of an interval matrix is computationally more difficult than the inverse of a real matrix. There are range of algorithms, which can calculate bounds on the interval hull, like interval Gaussian elimination (Neumaier 1990). Depending on the size and structure of the matrix computation of the interval hull is feasible.

For a linear regression in two variables the matrix $\mathbf{X}^T \mathbf{X}$ is a 2×2 matrix. Inverses for 2×2 matrix can be computed exactly using polytopes (Neumaier 1990). Even with the possibility of computing the inverse exactly there are several repeated variables in the expression for the least squares solution. Repeated variables can add additional inflation to the answer (Kreinovich *et al.* 2002). An algorithm which gives sharp interval bounds on the slope and intercept would be a welcome development.

For parameter bounding there exists a feasible algorithm for when the independent and dependent variables are uncertain (Jaulin *et al.* 2001).

4 Interpolation and extrapolation

So far our argument has been focused on estimating the coefficients for a linear equation. Given the interval hull solution, which are just bounds on the slope and the intercept for a line, straightforward interval computations can be used to estimate the value of the function. If the solution set is not rectangular then using interval arithmetic will give you conservative but non-optimal bounds on the range of values. If the solution set has been found using a polytope or a subpaving then the range can be found optimally. An algorithm called ImageSP is detailed in Jaulin *et al.* (2001) which when applied can evaluate the image of a function for particular value of the dependent variable.

5 Regression of fuzzy numbers

The only data requirement for regression on fuzzy numbers is that every fuzzy number have the same number of nested intervals. A fundamental property of interval arithmetic is that if $X \subseteq Z$, where X, Z are intervals, then $F(X) \subseteq F(Z)$, where F is a function composed of a finite number of binary interval operators. This property allows, for example, the posterior feasible set for parameter bounding to be computed efficiently.

The algorithm for parameter bounding based on SIVIA requires an initial search box that must enclose the solution set. Once a feasible posterior set for $\alpha = 0$, the widest interval in a fuzzy number, has been calculated the

interval hull for the set can be used as the initial search box for the next alpha level. This results for fuzzy numbers with decreasing width with increasing α is smaller and smaller search boxes as α goes to 1.

6 Overview of approaches

Two major approaches have been explored for fuzzy regression. The question is which of the two approaches is best. The answer whether to use fuzzy regression or interval least squares regression depends more on the particular dataset at hand than any intrinsic merits of the two methods. Dataset 1 which the error in the dependent variable are small interval least squares is the preferred approach. In many cases, parameter bounding would give the null set because there does not exist a single line which crosses through every datum. A null set is of little use to practitioner trying to make a prediction. When there is large uncertainty in the dependent variable then parameter bounding is more appropriate. When the posterior feasible set from parameter bounding is not null then interval least squares makes little sense. There is at least one solution which the sum of squares equals 0. Parameter bounding has an additional advantage in that it is easy to fit a nonlinear model as a linear model. Now there is no question whether the residuals meets the requirement of being normally distributed with a mean of 0.

References

- [1] Gay, D. M. Interval least squares—a diagnostic tool. In R. Moore, editor, *Reliability in Computing: The role of interval methods in scientific computing*, pages 183–205. Academic Press, New York, 1988.
- [2] Hargreaves, G. I. 2002. Interval analysis in MATLAB. *Numerical Analysis Report, No. 416*, Manchester Centre for Computational Mathematics, U.K.
- [3] Jaulin, L. and É. Walter. 1993. Guaranteed nonlinear parameter estimation via interval computations. *Interval Computations*, 3: 61–75.
- [4] Jaulin, L., M. Kieffer, O. Didrit, and É. Walter. 2001. *Applied Interval Analysis*. Springer-Verlag, London.
- [5] Kaufmann, A. and M. M. Gupta. 1991. *Introduction to Fuzzy Arithmetic Theory and Application*, Van Nostrand Reinhold, New York.

- [6] Kreinovich, V., L. Longpre, and J. J. Buckley. 2002. Are there efficient necessary and sufficient conditions for straightforward interval computations to be exact? In *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, pages 94–96, Toronto, Canada.
- [7] Marino, M. and F. Palumbo. 2003. Interval linear regression: an application to soil permeability analysis, in *Convegno intermedio della SIS*, Napoli.
- [8] Moore, R. 1992. Parameter sets for bounded-error data. *Mathematics and Computer and Information Science*, 34:113–119.
- [9] A. Neumaier. 1990. *Interval methods for systems of equations*. Cambridge University Press, Cambridge, U.K.
- [10] A. Neumaier. 2001. *Introduction to Numerical Analysis*. Cambridge University Press, Cambridge, U.K.
- [11] Norton, J. P. 1996. Rules for deterministic bounding in environmental modelling. *Ecological Modelling*, 86:157–161
- [12] Norton, J. P. Linear-model case, in bounds-based identification. 2002. In H. Unbehauen, editor, *Encyclopedia of Life Support Systems*, number 6.43.11. UNESCO, Oxford, UK.
- [13] Salia S. B. and S. Ferson. 1998. Fuzzy regression in fisheries science: some methods and applications. *Alaska Sea Grant College Program AK-SG-98-01*.
- [14] Schäfer, U. 2004. A linear complementarity problem with a P -Matrix. *SIAM Review*, 46(2):189–201.

7 Appendix

```
% Least squares regression for interval data (Dataset 1)
run /usr/local/matlab6/toolbox/intlab/startintlab

X = [1,1;1,3;1,5;1,7]
iX = intval(X)
iY = [infsup(20,21);infsup(30,31);infsup(25,26);infsup(33,34)]

iB = (iX' * iX)^-1 * iX' * iY
infsup(iB)
```