Interval Linear and Nonlinear Regression – New Paradigms, Implementations, and Experiments
or
New Ways of Thinking About Data Fitting

by
Jie Yang and
R. Baker Kearfott, rbk@louisiana.edu
Department of Mathematics, University of Louisiana at Lafayette

Credits: Arnold Neumaier, "Interval Linear Equations,", in *Interval Mathematics 1985, Lecture Notes in Computer Science 212*, ed. K. Nickel, Springer Verlag, 1985.

The Traditional Modus Operandi

- Classical regression involves fitting a model with a small number of parameters to a large number of data points.
- Typically least squares, minimax, or l_1 fits are used.
- The assumption is that the underlying model is exact, but there are errors in the data. The fits do not fit the data exactly, but minimize some metric of the distance from the data.

An Alternate Procedure

1. Start with a model, with a small number of parameters, that we assume can fit correct data exactly.

2. Perturb the data more and more until we can prove that the model admits an exact fit for some point data set within the perturbed data.

3. Output the point fit shown to be an exact fit to the perturbed data.

4. Also output bounds on the parameter values within which all exact solutions must lie, given the bounds on the data.

5. Output the bounds on the perturbed data.

A Third Possibility

1. Start with a model as above.
2. Instead of perturbing the data, *provide a priori* bounds within which the data is known to lie.
3. Try to prove that some point within the given bounds admits an exact solution.
4. *IF* existence of an exact solution can be proven, *THEN*
   (a) Output the parameter values corresponding to the exact solution.
   (b) Output bounds on the parameter values within which all exact solutions must lie, given the bounds on the data.

   ELSE Output "The model may not be appropriate for this data, or the assumptions on the bounds on the data may be incorrect."

An Alternate Procedure

Example

Suppose we suspect that $u = f(t) = x_1 t + x_2$ is a good model for the data

i	t	u
1	0	1
2	1	4
3	2	5
4	3	8

If the model fit the data exactly, then we would have

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \\ 5 \\ 8 \end{pmatrix}.$$

Example

(Continued)

- Due to "errors" in the data (say, in both the values of t and u), the system is inconsistent.

- However, suppose we have somehow determined error bounds for the data, so that the actual system of equations is contained in the interval system

$$\begin{pmatrix} [-0.09885, 0.09885] & [0.9012, 1.099] \\ [0.9012, 1.099] & [0.9012, 1.099] \\ [1.901, 2.099] & [0.9012, 1.099] \\ [2.901, 3.099] & [0.9012, 1.099] \end{pmatrix} x$$

$$= \begin{pmatrix} [0.7364, 1.264] \\ [3.736, 4.264] \\ [4.736, 5.264] \\ [7.736, 8.264] \end{pmatrix}.$$

Example

(Continued)

1. The above interval linear system $\boldsymbol{A}x = \boldsymbol{b}$ *can* have an exact solution in the sense that there are x and $A \in \boldsymbol{A}$, $b \in \boldsymbol{b}$ with $Ax = b$.

2. Beginning with the least squares solution $x = (2.2, 1.2)^T$ system representing the midpoints of the elements, we see that $\boldsymbol{A}x \cap \boldsymbol{b} \neq \emptyset$.

3. Due to a result of Beeck, this shows that x is in the solution set of $\boldsymbol{A}x = \boldsymbol{b}$.

4. Furthermore, following a procedure pointed out by Neumaier, the solution set of $\boldsymbol{A}x = \boldsymbol{b}$ is bounded by $\boldsymbol{x} = ([1.677, 2.813], [0.388, 2.273])^T$.

A Verification Principle

1. Consider $\boldsymbol{A}x = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{IR}^{m\times n}$, $\boldsymbol{b} \in \mathbb{IR}^m$, $m > n$.

2. Let Y be the Moore-Penrose pseudo-inverse of, say, the midpoint matrix of $\boldsymbol{A}$.

3. Any possible solutions of $\boldsymbol{A}x = \boldsymbol{b}$ must also be solutions of $Y\boldsymbol{A}x = Y\boldsymbol{b}$.

4. An interval Newton method can be applied to $Y\boldsymbol{A}x = Y\boldsymbol{b}$ to
   - determine non-existence, or
   - compute narrow(er) bounds on the solution set to $\boldsymbol{A}x = \boldsymbol{b}$.

Some Illustrations

Solution Sets to Overdetermined Interval Systems

Consider the system $\boldsymbol{A}x = \boldsymbol{b}$, where

$$\boldsymbol{A} = \begin{pmatrix} [0.9, 1.1] & 1 \\ [1.9, 2.1] & 1 \\ [2.9, 3.1] & 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} [1.9, 2.1] \\ [3.9, 4.1] \\ [5.9, 6.1] \end{pmatrix}.$$

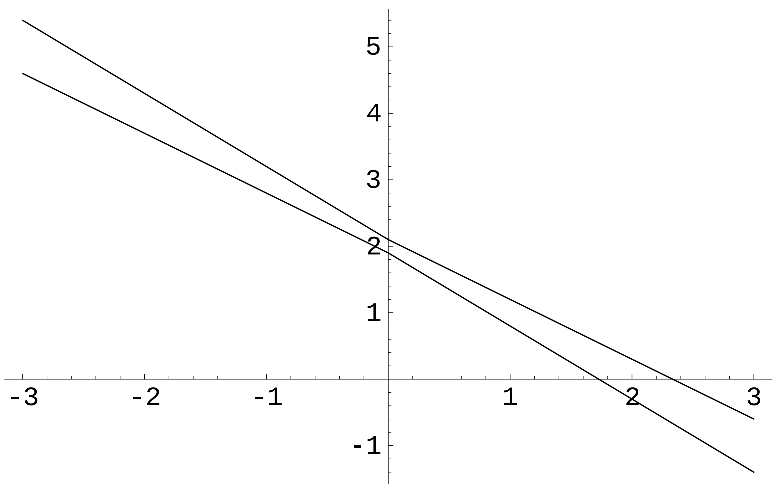
The first equation is satisfied where the following inequalities are true.

$$\begin{array}{l} -0.9x_1 + 1.9 \ \leq \ x_2 \ \leq \ -1.1x_1 + 2.1, \ x_1 \leq 0, \\ -1.1x_1 + 1.9 \ \leq \ x_2 \ \leq \ -0.9x_1 + 2.1, \ x_1 > 0. \end{array}$$

The graph of the solution of this set of inequalities appears on the next page.

Solution Set Graphs

The First Equation



$$[0.9, 1.1]x_1 + x_2 = [1.9, 2.1].$$

We combine this with the solution sets to

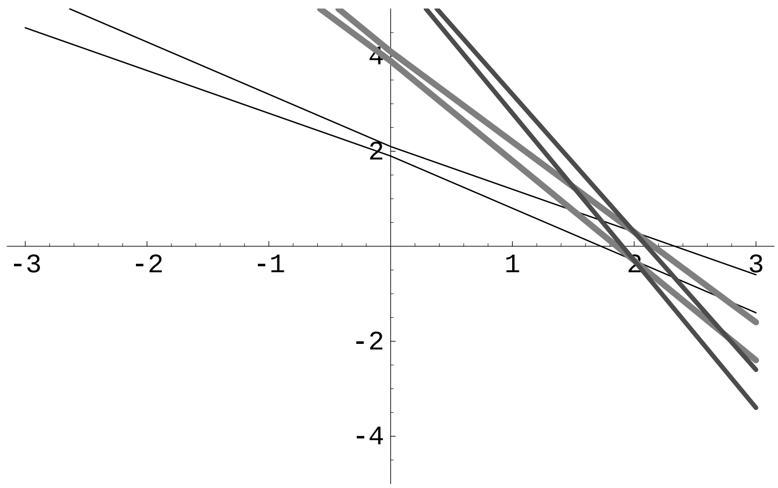
$$[1.9, 2.1]x_1 + x_2 = [3.9, 4.1]$$

and

$$[2.9, 3.1]x_1 + x_2 = [5.9, 6.1]$$

on the next graph.

Solution Set Graphs



The simultaneous solution set of

$$\begin{aligned} [0.9, 1.1]x_1 + x_2 &= [1.9, 2.1], \\ [1.9, 2.1]x_1 + x_2 &= [3.9, 4.1], \\ [2.9, 3.1]x_1 + x_2 &= [5.9, 6.1]. \end{aligned}$$

Relationships

- Traditional least squares:
  1. Compute the least squares solution to $Ax = b$.
  2. Define $\tilde{b} = b + r$, where r is the least squares residual $r = Ax - b$.
  3. If x and r were exact, then $Ax = \tilde{b}$ is satisfied exactly. Thus, $\boldsymbol{A}x \cap \square(b, \tilde{b}) \neq \emptyset$, where $\square(b, \tilde{b})$ is the smallest box containing b and $\tilde{b}$, and $\boldsymbol{A} = A$.
- The interval technique:
  - allows more flexibility (individual components in both b and A may be interactively perturbed);
  - gives *bounds* $\boldsymbol{x}$ on the solution set to $\boldsymbol{A}x = \boldsymbol{b}$, useful if data error bounds are known beforehand.
  - can show non-existence of solutions, with a-priori error bounds.