On the Tolerance Approach to Possibilistic Nonlinear Regression over Interval Data

Milan Hladík [1,2] Michal Černý [2]

[1] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

[2] Faculty of Computer Science and Statistics, University of Economics, Prague, Czech Republic

REC 2012, Brno, Czech Republic
June 13 – 15

Interval Linear Regression

Traditional linear regression model

$$y = X\theta + \varepsilon$$

where

- y ... vector of output data;
- X ... matrix of input data;
- θ ... vector of regression parameters;
- ε ... vector of disturbances.

Crisp input – interval output model

$$\text{“}\mathbf{y} = X\theta + \varepsilon\text{”}$$

Interval input – interval output model

$$\text{“}\mathbf{y} = \mathbf{X}\theta + \varepsilon\text{”}$$

Interval Linear Regression

Possibilistic solution concept

Determine an interval vector $\boldsymbol{\theta}$ such that each observation is "covered"

$$\mathbf{y}_i \subseteq \mathbf{X}_i\boldsymbol{\theta}, \quad i = 1, \dots, n.$$

- We want $\boldsymbol{\theta}$ as narrow as possible.
- This is a multi-criteria optimization problem!

Other solution concepts

- necessity solution $\dots \forall i = 1, \dots, n: \ \mathbf{y}_i \supseteq X_i\boldsymbol{\theta}$.
- weak possibilistic solution $\dots \forall i = 1, \dots, n: \ \mathbf{y}_i \cap \mathbf{X}_i\boldsymbol{\theta} \neq \emptyset$.
- strong possibilistic solution $\dots \forall i = 1, \dots, n: \ \forall X_i' \in \mathbf{X}_i: \ \mathbf{y}_i \subseteq X_i'\boldsymbol{\theta}$.

Interval Linear Regression

The tolerance approach

Consider $\boldsymbol{\theta}$ in the form of $\boldsymbol{\theta} = [\theta^c - \delta^* \cdot c, \theta^c + \delta^* \cdot c]$, where

- θ^c is determined from (y^c, X^c) using e.g. least squares;
- c is given by a user, e.g.
  - $c = |\theta^c|$ ... relative tolerance,
  - $c = (1, \dots, 1)^T$ ... absolute tolerance,
  - if $c_i = 0$ then θ_i is forced to be crisp
- δ^* is defined as

$$\delta^* = \inf\{\delta \geq 0;\ \mathbf{y}_i \subseteq \mathbf{X}_i[\theta^c - \delta \cdot c, \theta^c + \delta \cdot c],\ i = 1, \dots, n\}.$$

Properties of δ^*

- it is easy to calculte, e.g. for the crisp input – ouput model

$$\delta^* = \max_{i:|X|_i c > 0} |y_i - X_i \theta^c| \,/\, (|X|_i c)$$

- its interpretation: For relative tolerances, it suffices to perturb θ^c by no more than $100 \cdot \delta^*\%$ to cover all data
- it can serve as a goodness-of-fit measure of the model

Interval Nonlinear Regression

From now on, we restrict ourselves to *crisp input – interval output models*. That is, X is crisp, $\mathbf{y}$ is interval. (The ideas can be extended further.)

Model

Consider a (nonlinear) regression function

$$f(x; \theta_1, \dots, \theta_p),$$

where

- x is a *data variable* (possibly a vector of variables),
- $\theta = (\theta_1, \dots, \theta_p)^T$ are *regression parameters*.

Possibilistic solution

In the crisp input – interval output regression model

$$\text{“}\mathbf{y}_i = f(x_i, \boldsymbol{\theta})\text{”}$$

we seek for a narrow interval vector $\boldsymbol{\theta}$ such that

$$\mathbf{y}_i \subseteq f(x_i, \boldsymbol{\theta}), \quad i = 1, \dots, n.$$

The tolerance approach: analogy of the linear case

Method

- *Step 1.* Determine θ^c using traditional nonlinear regression methods for data (X, y^c).
- *Step 2.* Choose c. [For example: choose $c = |\theta^c|$ for relative tolerances.]
- *Step 3.* Find the minimal tolerance quotient δ^* such that

$$\mathbf{y}_i \subseteq f(x_i, \boldsymbol{\theta}^*), \quad i = 1, \ldots, n, \tag{1}$$

where

$$\boldsymbol{\theta}^* = [\theta^c - \delta^* \cdot c, \ \ \theta^c + \delta^* \cdot c].$$

Now, $\boldsymbol{\theta}^*$ is the resulting interval vector of regression parameters.

Observation

If the inclusion (1) can be tested efficiently, then δ^* can be found using binary search (under certain assumptions).

We need to determine $f(x, \boldsymbol{\theta})$...

Theorem (Gaganov, 1981)

In general, computing $f(x, \boldsymbol{\theta})$ is NP-hard (even its ε-approximation).

Definition ($\mathcal{A}$-type nonlinear regression models)

The function $f(x, \theta)$ can be expressed as a formula

- containing $+, -, \times, \div$;
- elementary functions which are easy-to-evaluate over intervals (e.g. exp, log, etc.);
- is monotone with respect to each parameter $\theta_1, \ldots, \theta_p$ which appears *more than once* in the formula.

- For $\mathcal{A}$-type model, interval arithmetic calculates exactly $f(x, \boldsymbol{\theta})$.
- Otherwise, it may overestimate. In general, checking exactness of interval arithmetic is NP-hard (Kreinovich, Longpré, Buckley, 2003), but we may sometimes utilize endpoint analysis.

Examples of $\mathcal{A}$-type models

- The logistic growth model

$$f(x;\theta_1,\theta_2,\theta_3)=\frac{\theta_1}{1+e^{-\theta_2(x-\theta_3)}};$$

- Gompertz growth model

$$f(x;\theta_1,\theta_2,\theta_3)=\theta_1\cdot e^{-e^{-\theta_2(x-\theta_3)}};$$

- estimation of the degree of a polynomial in the form

$$f(x;\theta_1,\theta_2,\theta_3)=\theta_1+\theta_2 x+\theta_3 x^{\theta_4};$$

- Berry's model (used in agriculture for modeling the crop yield as a function of the density of planting)

$$f(x_1,x_2;\theta_1,\theta_2,\theta_3,\theta_4)=\left(\theta_1+\theta_2\left(\frac{1}{x_1}+\frac{1}{x_2}\right)+\frac{\theta_3}{x_1x_2}\right)^{-\theta_4};$$

- oscillation model

$$f(x;\theta_1,\theta_2,\theta_3)=\theta_1 e^{-\theta_2 x}\cos(\theta_3 x).$$

Example 1: degradation of material

- We measure the degree of degradation of a material (y) as a function of time (x) for which the material is exposed to unfavorable conditions (such as temperature or pressure).
- The degree of disruption is measured on a discrete scale $0, \dots, 10$, where 0 means "no damage", 1 means "very mild damage", ..., and 10 means "totally damaged".
- The values of y are determined by experts (say, by visual inspection of constructions where the material has been used).

Due to a certain subjectivity of experts, it is appropriate to consider the grade $y \in \{1, \dots, 9\}$ as an interval, say of the form

$$[\underline{y}, \overline{y}] = [y - 0.5, y + 0.5].$$

We model the dependence of y on x using the Gompertz curve

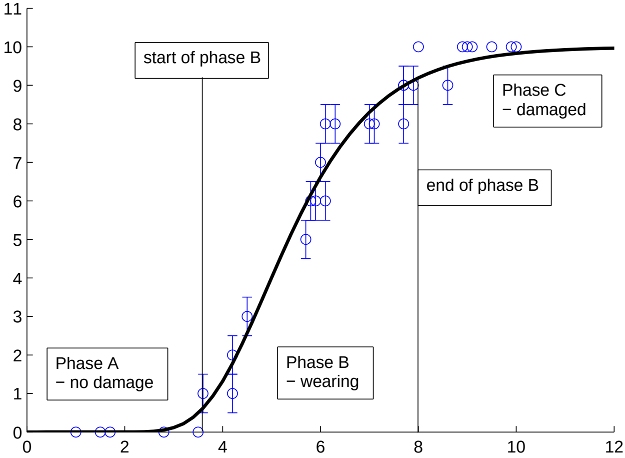
$$y = 10e^{-e^{-\theta_1(x-\theta_2)}}.$$

This is an $\mathcal{A}$-class model.

Example 1 continued

First, we fit the centers of data using nonlinear least squares, resulting in the estimated parameters

$$\widehat{\theta}_1 = 0.795, \quad \widehat{\theta}_2 = 4.887.$$

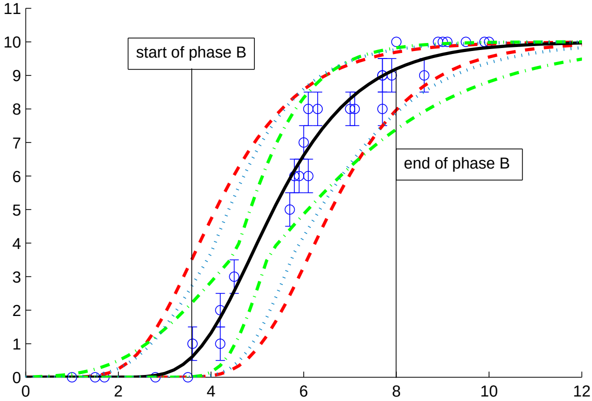


Example 1 continued

- Now we would like to extend the estimated crisp values $\theta_1^c := \widehat{\theta}_1$ and $\theta_2^c := \widehat{\theta}_2$ to interval values covering all observations.
- We observe that the points $y \in \{0, 10\}$ can never be covered with the Gompertz curve. From now on, we restrict ourselves to the data in the B-phase only.
- **Choice 1.** We set $c = \binom{0.795}{4.887}$ (i.e., relative tolerances). We get the value $\delta^* = 0.183$. The data are covered by the intervals $[(1 - 0.183) \cdot 0.795, (1 + 0.183) \cdot 0.795]$ and $[(1 - 0.183) \cdot 4.887, (1 + 0.183) \cdot 4.887]$ for θ_1 and θ_2, respectively. We conclude that *it suffices to perturb the values* $\widehat{\theta}_1, \widehat{\theta}_2$ *by no more than* 18.3% *in order all intervals be covered.*
- **Choice 2.** We set $c = \binom{1}{1}$ (i.e., absolute tolerances). We get the value $\delta^* = 0.360$. The data are covered by the intervals $[0.795 - 0.36, 0.795 + 0.36]$ and $[4.887 - 0.36, 4.887 + 0.36]$ for θ_1 and θ_2, respectively. We conclude that *it suffices to perturb the values* $\widehat{\theta}_1, \widehat{\theta}_2$ *by no more than* 0.36 *in order all intervals be covered.*

Example 1 continued

- **Choice 3.** We set $c = \binom{0}{4.887}$. This models the situation that the dynamics of degradation is kept constant and we can perturb only the shift θ_2 of the Gompertz curve to cover the data. We get the value $\delta^* = 0.254$. The data are covered by the interval $[(1-0.254)\cdot 4.887, (1+0.254)\cdot 4.887]$ for θ_2.



Example 2 — change-in-phase regression

- In this example we consider a *non-$\mathcal{A}$-type model*.
- We will evaluate the non-$\mathcal{A}$-type regression function $f(x;\boldsymbol{\theta})$ using interval arithmetic. Then, we can face *redundancy*.

Assume that the dependent variable y depends on the explanatory variable x according to the simple linear relationship

$$y = \theta_1 + \theta_2 x.$$

We know that *the relationship changes in an unknown point*. Assume that the point of change is continuous and smooth. It is suitable to use a model of the form

$$y = (1 - S(x))(\theta_1 + \theta_2 x) + S(x)(\theta_3 + \theta_4 x),$$

where S is a suitable nondecreasing function with $S(\mathbb{R}) = (0, 1)$.

$S(x)$ is called a “switching function”. We will use the logistic function

$$L(x;\theta_5,\theta_6) = \frac{1}{1 + e^{-\theta_5(x-\theta_6)}}.$$

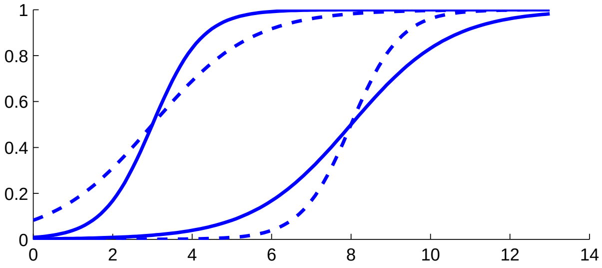
Example 2 continued

In the model

$$y = (1 - S(x))(\theta_1 + \theta_2 x) + S(x)(\theta_3 + \theta_4 x),$$

the switching function $S(x) = L(x; \theta_5, \theta_6)$ plays the following role:

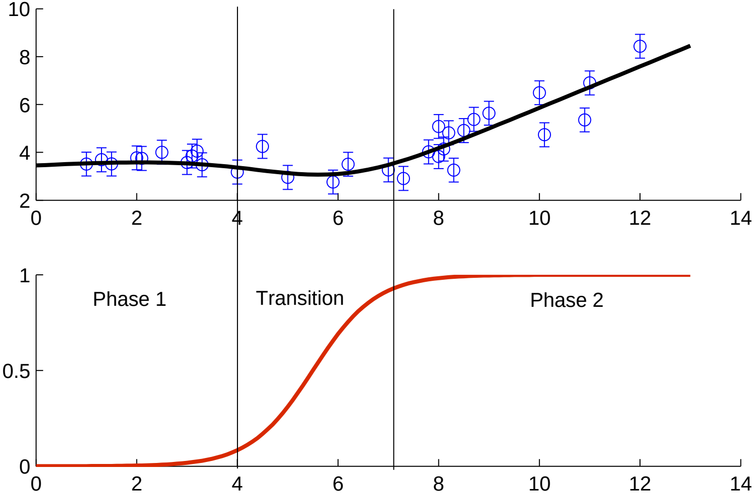
- when $S(x) \approx 0$, then the data follow the model $\theta_1 + \theta_2 x$ ("Phase 1");
- when $S(x) \approx 1$, then the data follow the model $\theta_3 + \theta_4 x$ ("Phase 2");
- when $0 \ll S(x) \ll 1$, then the data are in a "transition phase" between Phase 1 and Phase 2.



Example 2 continued

Using nonlinear least squares for centers of data we get

$$\widehat{\theta}_1 = 3.47, \quad \widehat{\theta}_2 = 0.12, \quad \widehat{\theta}_3 = -2.73, \quad \widehat{\theta}_4 = 0.86, \quad \widehat{\theta}_5 = 1.60, \quad \widehat{\theta}_6 = 5.50.$$



Example 2 continued

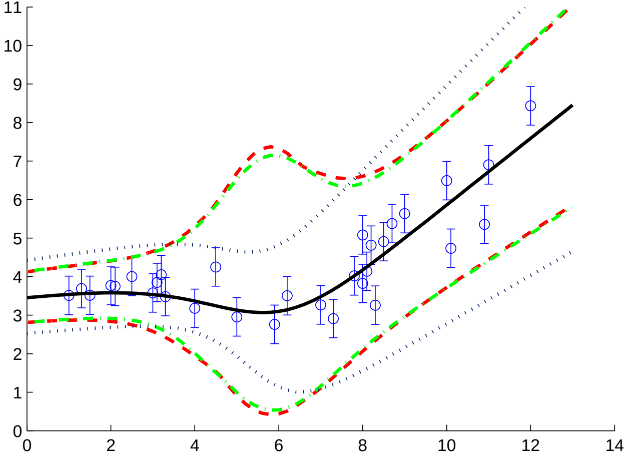
We apply the tolerance approach for calculation of the interval regression parameters.

We set $\theta^c = (3.47, 0.12, -2.73, 0.86, 1.60, 5.50)^T$ and we consider three choices of c:

- **Choice 1:** $c = |\theta^c| = (3.47, 0.12, 2.73, 0.86, 1.60, 5.50)^T$ (relative tolerances), with the resulting value $\delta^* = 0.186$. *It suffices to perturb the regression coefficients θ^c by no more that 18.6% in order all data be covered.*
- **Choice 2:** $c = (3.47, 0.12, 2.73, 0.86, 0, 5.50)^T$ (relative tolerances assuming that the dynamics of the transition phase is fixed), with the resulting value $\delta^* = 0.190$;
- **Choice 3:** $c = (3.47, 0.12, 2.73, 0.86, 0, 0)^T$ (relative tolerances assuming that the dynamics and location of the transition phase is fixed), with the resulting value $\delta^* = 0.273$.

Example 2 continued

We plot the enclosures as a function of x, where the expression $f(x; [\theta^c - \delta^* \cdot c, \theta^c + \delta^* \cdot c])$ was evaluated using interval arithmetic. (Observe that it can be redundant since this is not an $\mathcal{A}$-type model.)



Conclusion and future work

- For a non-$\mathcal{A}$-type model, the computed δ^* is a lower bound on the optimum.
  An upper bound can be computed by the overestimation formula for the mean value or slope enclosures.
- For interval input – interval/crisp output model, it is also natural to consider the solution concept

$$\mathbf{y}_i \subseteq \left[\max_{X_i \in \mathbf{X}_i} \underline{f}(X_i, \boldsymbol{\theta}), \min_{X_i \in \mathbf{X}_i} \overline{f}(X_i, \boldsymbol{\theta}) \right], \quad i = 1, \ldots, n$$

  instead of

$$\mathbf{y}_i \subseteq f(\mathbf{X}_i, \boldsymbol{\theta}), \quad i = 1, \ldots, n.$$

  It leads to (Kaucher) extended interval arithmetic.