Towards More Realistic Interval Models in Econometrics

Songsak Sriboonchitta\textsuperscript{1}, Thach N. Nguyen\textsuperscript{2}, Olga Kosheleva\textsuperscript{3}, and Vladik Kreinovich\textsuperscript{3}

\textsuperscript{1} Chiang Mai University, Chiang Mai, Thailand
\textsuperscript{2} Banking University, Ho Chi Minh City, Vietnam
\textsuperscript{3} University of Texas at El Paso, El Paso, Texas, USA
500 W. University, El Paso, TX 79968, USA

songsakecon@gmail.com, Thachnn@buh.edu.vn, olgak@utep.edu, vladik@utep.edu

\textbf{What is econometrics.} The main objective of econometrics is to use the past economic data to predict – and, if needed, change – the future economic behavior. The simplest – and often efficient – way to predict is to find a linear dependence between the desired future value $y$ and the present and past values $x_1, \ldots, x_n$ of this and related quantities:

$$y \approx c_0 + \sum_{i=1}^{n} c_i \cdot x_i$$

for some coefficients $c_i$. The coefficients $c_i$ can be determined from the available data $(x_1^{(k)}, \ldots, x_n^{(k)}, y^{(k)}), 1 \leq k \leq K$, by using the Least Squares method

$$\sum_{k=1}^{K} \left( y^{(k)} - \left( c_0 + \sum_{i=1}^{n} c_i \cdot x_i^{(k)} \right) \right)^2,$$  \hspace{1cm} (1)

if the approximation error is normally distributed – or, in general, by other particular cases of the Maximum Likelihood method.

In the case of Least Squares, differentiation leads to an easy-to-solve system of linear equations for the unknowns $c_i$.

\textbf{Why interval models in econometrics.} For stock trading, we have millions of records daily, corresponding to seconds and even milliseconds. A few decades ago, it was not possible to process all this data. So, econometricians considered only one value per day – e.g., the price at the end of the working day.

Nowadays, with more computational power at our disposal, we can consider many more data points. Practitioners expected that by taking into account more price values per day – i.e., more data – we can get better predictions. Somewhat surprisingly, it turned out that predictions got worse. This means that most daily price fluctuations are irrelevant for predictions, they are noise that only make prediction worse.
The same thing happened if instead of the price $x_i$ at the end of the working day, practitioners considered two numbers: the smallest price $\underline{x}_i$ during the day and the largest price $\overline{x}_i$ during the day. The only idea that helped improve the prediction accuracy was replacing the previous value $x_i$ with some more relevant value from the corresponding interval $[\underline{x}_i, \overline{x}_i]$.

**How interval data is treated now.** In situations when instead of the exact values $x_i^{(k)}$ and $y^{(k)}$, we only know intervals $[\underline{x}_i^{(k)}, \overline{x}_i^{(k)}]$ and $[\underline{y}^{(k)}, \overline{y}^{(k)}]$, researchers proposed to use the values $y^{(k)} = \alpha \cdot \overline{y}^{(k)} + (1 - \alpha) \cdot \underline{y}^{(k)}$ and $x_i^{(k)} = \alpha \cdot \overline{x}_i^{(k)} + (1 - \alpha) \cdot \underline{x}_i^{(k)}$, for some special value $\alpha$ – usually, $\alpha = 0$, $\alpha = 0.5$, or $\alpha = 1$. This led to some improvement in prediction accuracy.

Even better results were obtained when they tried, instead of fixing a value $\alpha$, to find the value $\alpha$ for which the mean squared error is the smallest (or, more generally, the Maximum Likelihood is the largest).

The optimization problem is no longer quadratic, but it is quadratic with respect to $c_i$ and with respect to $\alpha$, so we can solve it by inter-changingly minimizing over $c_i$ and over $\alpha$.

**Discussion.** As we have mentioned, deviations from the typical daily value are random. One day, we have increases, another day, decreases. So, instead of fixing the same midpoint (for $\alpha = 0.5$) or, more generally, the same $\alpha$-point for all $i$ and $k$, it makes more sense to select possibly different points from different intervals, i.e., to select values $c_i$, $x_i^{(k)} \in [\underline{x}_i^{(k)}, \overline{x}_i^{(k)}]$, and $y^{(k)} \in [\underline{y}^{(k)}, \overline{y}^{(k)}]$ that minimize the expression (1).

It may seem that the above $\alpha$-approach is a good first approximation for this optimization problem. We usually take $\alpha \in (0, 1)$, thus $x_i^{(k)} \in (\underline{x}_i^{(k)}, \overline{x}_i^{(k)})$, and $y^{(k)} \in (\underline{y}^{(k)}, \overline{y}^{(k)})$. So, it seems like in our problem, we should also look for a minimum inside the corresponding intervals.

But in this case, the derivatives of (1) with respect to $y^{(k)}$ and $x_i^{(k)}$ would be equal to 0, and thus, for all $k$, we would have exact equality $y^{(k)} = c_0 + \sum_{k=1}^{K} c_i \cdot x_i^{(k)}$. In most practical problems, it is not possible to fit all the available intervals with the exact dependence. This means that in the optimal solution, for some $k$, we are not inside the intervals, we are at an endpoint.

**How to actually solve the corresponding optimization problem.** Similarly to the $\alpha$-approach, we can perform iterative optimization. Specifically, we start, e.g., with midpoints $\overline{y}^{(k)}$ and $\underline{x}_i^{(k)}$. Then, inter-changingly, we either find $c_i$ (while keeping $y^{(k)}$ and $x_i^{(k)}$ fixed), or keep $c_i$ fixed and find $x_i^{(k)}$ and $y^{(k)}$ from the corresponding intervals (by solving the corresponding feasible-to-solve convex constraint optimization problem).