Towards More Realistic Interval Models in Econometrics

Songsak Sriboonchitta[1], Thach N. Nguyen[2],
Olga Kosheleva[3], and Vladik Kreinovich[3]

[1]Chiang Mai University, Chiang Mai, Thailand
[2]Banking University, Ho Chi Minh City, Vietnam
[3]University of Texas at El Paso, El Paso, TX 79968, USA
songsakecon@gmail.com, Thachnn@buh.edu.vn,
olgak@utep.edu, vladik@utep.edu

1. Why Interval Models In General: A Brief Reminder

- In most application areas, values of the quantities come from measurements.
- Measurement of a physical quantity x is always approximate:
  - it produces a value $\widetilde{x}$
  - which is, in general, different from the desired actual value x.
- In many case, we have no information about the probabilities of different values of the measurement error

$$\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x.$$

- We only know the upper bound Δ on the absolute value of the measurement error: $|\Delta x| \le \Delta$.

2. Why Interval Models In General (cont-d)

- Often, we only know the upper bound Δ on the absolute value of the measurement error: $|\Delta x| \le \Delta$.
- In such cases:
  - once we know the measurement result $\widetilde{x}$,
  - the only information that we have about the actual (unknown) value x is that $x \in [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.

3. What Is Econometrics: A Brief Reminder

- In econometrics – a quantitative study of economics – we deal with values like prices, indexes, etc.
- Most of these values are known exactly, there is no measurement uncertainty.
- The stock's prices, the amounts of stocks traded – all these numbers are known exactly.
- So, at first glance, there seems to be no need for interval models in econometrics.
- But, as we will show, there is such a need.
- Indeed, the main objective of econometrics is:
  - to use the past economic data
  - to predict – and, if needed, change – the future economic behavior.

4. What Is Econometrics (cont-d)

- 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:

$$y \approx c_0 + \sum_{i=1}^{n} c_i \cdot x_i \text{ for some coefficients } c_i.$$

- The coefficients can be determined from the available data $\left(x_1^{(k)}, \ldots, x_n^{(k)}, y^{(k)}\right)$, $1 \le k \le K$.
- When the approximation error is normally distributed, we can use the Least Squares method

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

5. What Is Econometrics (cont-d)

- In general, we can use 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 c_i.
- 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:
  - instead of considering all second-by-second prices of a stock,
  - 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.

6. Why Interval Models in Econometrics

- Practitioners expected that:
  - by taking into account more price values per day – i.e., more data,
  - then can get better predictions.
- Somewhat surprisingly, it turned out that predictions got worse.
- Namely, it turned out that most daily price fluctuations are irrelevant for prediction purposes.
- They constitute noise whose addition only makes the prediction worse.
- The same thing happened if instead of a single value x_i, practitioners considered *two* numbers:
  - the smallest price $\underline{x}_i$ during the day and
  - the largest price $\overline{x}_i$ during the day.

7. Interval Models in Econometrics (cont-d)

- Many attempts to use extra data only made predictions worse.
- 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 $[\underline{x}_i, \overline{x}_i]$.

8. How Interval Data Is Treated Now

- We consider situations when:
  - instead of the exact values $x_i^{(k)}$ and $y^{(k)}$,
  - we only know intervals $\left[\underline{x}_i^{(k)}, \overline{x}_i^{(k)}\right]$ and $\left[\underline{y}^{(k)}, \overline{y}^{(k)}\right]$.
- To deal with such situations, researchers proposed to use the values

$$y^{(k)} = \alpha \cdot \overline{y}^{(k)} + (1 - \alpha) \cdot \underline{y}^{(k)} \text{ and}$$

$$x_i^{(k)} = \alpha \cdot \overline{x}_i^{(k)} + (1 - \alpha) \cdot \underline{x}_i^{(k)}.$$

- Here, α is some special value – usually, $\alpha = 0$, $\alpha = 0.5$, or $\alpha = 1$.
- This lead to some improvement in prediction accuracy.

9. How Interval Data Is Treated Now (cont-d)

- Even better results were obtained when they tried:
  - instead of fixing a value α,
  - to find the value α 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.
- However, it is quadratic with respect to c_i and with respect to α.
- So we can solve it by inter-changingly:
  - minimizing over c_i and
  - minimizing over α.

10. Discussion

- Deviations from the typical daily value are random.
- One day, they are mostly increasing, another day, they are mostly decreasing, so:
  - instead of fixing the same α for all i and k,
  - it makes more sense to select possibly different points from different intervals,
  - i.e., to select values c_i, $x_i^{(k)} \in \left[\underline{x}_i^{(k)}, \overline{x}_i^{(k)}\right]$, and $y^{(k)} \in \left[\underline{y}^{(k)}, \overline{y}^{(k)}\right]$ that minimize the expression

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

- It may seem that the existing α-approach is a good first approximation for this optimization problem.
- However, in the α-approach, we usually take $\alpha \in (0, 1)$.

11. Discussion (cont-d)

- When $\alpha \in (0,1)$, we always have $x_i^{(k)} \in \left(\underline{x}_i^{(k)}, \overline{x}_i^{(k)}\right)$, and $y^{(k)} \in \left(\underline{y}^{(k)}, \overline{y}^{(k)}\right)$.
- So, the minimum is attained inside the corresponding intervals.
- Thus, it seems like in our problem, we should also look for a minimum inside the corresponding intervals.
- But then, the derivatives of the objective function with respect to $y^{(k)}$ and $x_i^{(k)}$ would be equal to 0.
- 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.

12. Discussion (cont-d)

- If we were always inside the corresponding intervals, then we would always have equalities

$$y^{(k)} = c_0 + \sum_{k=1}^{K} c_i \cdot x_i^{(k)}.$$

- However, often, we cannot have all equalities.
- This means that in the optimal solution, for some k:
  - we are not inside the intervals,
  - we reach the endpoints of some of the given intervals.

13. How to Actually Solve the Corresponding Optimization Problem

- Similarly to the α-approach, we can perform iterative optimization.
- Specifically, we start, e.g., with midpoints $y^{(k)}$ and $x_i^{(k)}$.
- Then inter-changingly:
  - we find c_i (while keeping $y^{(k)}$ and $x_i^{(k)}$ fixed), and
  - we keep c_i fixed and find $x_i^{(k)}$ and $y^{(k)}$ from the corresponding intervals.
- On each step, we get a feasible-to-solve convex constraint optimization problem.

14. Acknowledgments

This work was supported in part:

- by the Department of Economics, Chiang Mai University, Thailand, and
- by the US National Science Foundation grant HRD-1242122 (Cyber-ShARE Center).