Why Threshold Models: A Theoretical Explanation

Thongchai Dumrongpokaphan, Vladik Kreinovich, and Songsak Sriboonchitta

Abstract Many economic phenomena are well described by linear models. In such models, the predicted value of the desired quantity – e.g., the future value of an economic characteristic – linearly depends on the current values of this and related economic characteristic and on the numerical values of external effects. Linear models have a clear economic interpretation: they correspond to situations when the overall effect does not depend, e.g., on whether we consider a loose federation as a single country or as several countries. While linear models are often reasonably accurate, to get more accurate predictions, we need to take into account that real-life processes are nonlinear. To take this nonlinearity into account, economists use piece-wise linear (*threshold*) models, in which we have several different linear dependencies in different domains. Surprisingly, such piece-wise linear models often work better than more traditional models of non-linearity – e.g., models that take quadratic terms into account. In this paper, we provide a theoretical explanation for this empirical success.

1 Formulation of the Problem

Linear models are often successful in econometrics. In econometrics, often, linear models are efficient, when the values $q_{1,t}, \ldots, q_{k,t}$ of quantities of interest $q_1, \ldots, q_k$

Thongchai Dumrongpokaphan
Department of Mathematics, Faculty of Science, Chiang Mai University, Thailand
e-mail: tcd43@hotmail.com

Vladik Kreinovich
University of Texas at El Paso, El Paso, TX 79968, USA
e-mail: vladik@utep.edu

Songsak Sriboonchitta
Faculty of Economics, Chiang Mai University, Thailand
e-mail: songsakecon@gmail.com

at time t can be predicted as linear functions of the values of these quantities at previous moments of time $t-1$, $t-2$, ..., and of the current (and past) values $e_{m,t}, e_{m,t-1}, \ldots$ of the external quantities $e_1, \ldots, e_n$ that can influence the values of the desired characteristics:

$$q_{i,t} = a_i + \sum_{j=1}^{k} \sum_{\ell=1}^{\ell_0} a_{i,j,\ell} \cdot q_{j,t-\ell} + \sum_{m=1}^{n} \sum_{\ell=0}^{\ell_0} b_{i,m,\ell} \cdot e_{m,t-\ell}; \tag{1}$$

see, e.g., [3, 4, 7] and references therein.

At first glance, this ubiquity of linear models is in line with general ubiquity of linear models in science and engineering. At first glance, the ubiquity of linear models in econometrics is not surprising, since linear models are ubiquitous in science and engineering in general; see, e.g., [5].

Indeed, we can start with a general dependence

$$q_{i,t} = f_i\left(q_{1,t}, q_{1,t-1}, \ldots, q_{k,t-\ell_0}, e_{1,t}, e_{1,t-1}, \ldots, e_{n,t-\ell_0}\right). \tag{2}$$

In science and engineering, the dependencies are usually smooth [5]. Thus, we can expand the dependence in Taylor series and keep the first few terms in this expansion. In particular, in the first approximation, when we only keep linear terms, we get a linear model.

Linear models in econometrics are applicable way beyond the Taylor series explanation. In science and engineering, linear models are effective in a small vicinity of each state, when the deviations from a given state are small and we can therefore safely ignore terms which are quadratic (or of higher order) in terms of these deviations.

However, in econometrics, linear models are effectively even when deviations are large and quadratic terms cannot be easily ignored; see, e.g., [3, 4, 7]. How can we explain this unexpected efficiency?

Why linear models are ubiquitous in econometrics. A possible explanation for the ubiquity of linear models in econometrics was proposed in [7]. Let us illustrate this explanation on the example of formulas for predicting how the country's Gross Domestic Product (GDP) $q_{1,t}$ changes with time t. To estimate the current year's GDP, it is reasonable to use:

- GDP values in the past years, and
- different characteristics that affect the GDP, such as the population size, the amount of trade, the amount of minerals extracted in a given year, etc.

In many cases, the corresponding description is un-ambiguous. However, in many other cases, there is an ambiguity in what to consider a country. Indeed, in many cases, countries form a loose federation: European Union is a good example. Most of European countries have the same currency, there are no barriers for trade and for movement of people between different countries, so, from the economic viewpoint, it make sense to treat the European Union as a single country. On the

other hand, there are still differences between individual members of the European Union, so it is also beneficial to view each country from the European Union on its own.

Thus, we have two possible approaches to predicting the European Union's GDP:

- we can treat the whole European Union as a single country, and apply the formula (2) to make the desired prediction;
- alternatively, we can apply the general formula (2) to each country $c = 1, \ldots, C$ independently

$$q_{i,t}^{(c)} = f_i\left(q_{1,t}^{(c)}, q_{1,t-1}^{(c)}, \ldots, q_{k,t-\ell_0}^{(c)}, e_{1,t}^{(c)}, e_{1,t-1}^{(c)}, \ldots, e_{n,t-\ell_0}^{(c)}\right). \tag{3}$$

  and then add up the resulting predictions.

The overall GDP $q_{1,t}$ is the sum of GDPs of all the countries:

$$q_{1,t} = q_{1,t}^{(1)} + \ldots + q_{1,t}^{(C)}.$$

Similarly, the overall population, the overall trade, etc., can be computed as the sum of the values corresponding to individual countries:

$$e_{m,t} = e_{m,t}^{(1)} + \ldots + e_{m,t}^{(C)}.$$

Thus, the prediction of $q_{1,t}$ based on applying the formula (2) to the whole European Union takes the form

$$f_i\left(q_{1,t}^{(1)} + \ldots + q_{1,t}^{(C)}, \ldots, e_{n,t-\ell_0}^{(1)} + \ldots + e_{n,t-\ell_0}^{(C)}\right),$$

while the sum of individual predictions takes the form

$$f_i\left(q_{1,t}^{(1)}, \ldots, e_{n,t-\ell_0}^{(1)}\right) + \ldots + f_i\left(q_{1,t}^{(C)}, \ldots, e_{n,t-\ell_0}^{(C)}\right).$$

Thus, the requirement that these two predictions return the same result means that

$$f_i\left(q_{1,t}^{(1)} + \ldots + q_{1,t}^{(C)}, \ldots, e_{n,t-\ell_0}^{(1)} + \ldots + e_{n,t-\ell_0}^{(C)}\right) =$$

$$f_i\left(q_{1,t}^{(1)}, \ldots, e_{n,t-\ell_0}^{(1)}\right) + \ldots + f_i\left(q_{1,t}^{(C)}, \ldots, e_{n,t-\ell_0}^{(C)}\right).$$

In mathematical terms, this means that the function f_i should be *additive*.

It also makes sense to require that very small changes in q_i and e_m lead to small changes in the predictions, i.e., that the function f_i be continuous. It is known that every continuous additive function is linear (see, e.g., [1]) – thus the above requirement explains the ubiquity of linear econometric models.

Need to go beyond linear models. While linear models are reasonably accurate, the actual econometric processes are often non-linear. Thus, to get more accurate predictions, we need to go beyond linear models.

A seemingly natural idea: take quadratic terms into account. As we have mentioned earlier, linear models correspond to the case when we expand the original dependence in Taylor series and keep only linear terms in this expansion. From this viewpoint, if we want to get a more accurate model, a natural idea is to take into account next order terms in the Taylor expansion – i.e., quadratic terms.

The above seemingly natural idea works well in science and engineering, but in econometrics, threshold models are often better. Quadratic models are indeed very helpful in science and engineering [5]. However, surprisingly, in econometrics, different types of models turn out to be more empirically successful: namely, so-called *threshold models* in which the expression f_i in the formula (2) is piece-wise linear; see, e.g., [2, 6, 8, 9, 10].

Terminological comment. Piece-wise linear models are called *threshold models* since in the simplest case of a dependence on a single variable $q_{1,t} = f_1(q_{1,t-1})$, such models can be described by listing:

- thresholds $T_0 = 0, T_1, \ldots, T_S, T_{S+1} = \infty$ separating different linear expressions, and
- linear expressions corresponding to each of the intervals $[0, T_1]$, $[T_1, T_2]$, $\ldots$, $[T_{S-1}, T_S]$, $[T_S, \infty)$:

$$q_{1,t} = a^{(s)} + a_1^{(s)} \cdot q_{1,t-1} \text{ when } T_s \le q_{1,t-1} \le T_{s+1}.$$

Problem and what we do in this paper. The challenge is how to explain the surprising efficiency of partial-linear models in econometrics.

In this paper, we provide such an explanation.

2 Our Explanation

Main assumption behind linear models: reminder. As we have mentioned in the previous section, the ubiquity of linear models can be explained if we assume that for loose federations, we get the same results whether we consider the whole federation as a single country or whether we view it as several separate countries.

A similar assumption can be made if we have a company consisting of several reasonable independent parts, etc.

This assumption needs to be made more realistic. If we always require the above assumption, then we get exactly linear models. The fact that in practice, we encounter some non-linearities means that the above assumption is not always satisfied.

Thus, to take into account non-linearities, we need to replace the above too-strong assumption with a more realistic one.

How can we make the above assumption more realistic: analysis of the problem. It should not matter that much if inside a loose federation, we move an area from

one country to another – so that one becomes slightly bigger and another slightly smaller – as long as the overall economy remains the same.

However, from the economic sense, it makes sense to expect somewhat different results from a "solid" country – in which the economics is tightly connected – and a loose federation of sub-countries, in which there is a clear separation between different regions. Thus:

- instead of requiring that the results of applying (2) to the whole country lead to the same prediction as results of applying (2) to sub-countries,
- we make a weaker requirement: that the sum of the result of applying (2) to sub-countries should not change if we slightly change the values within each sub-country – as long as the sum remains the same.

The crucial word here is "slightly". There is a difference between a loose federation of several economies of about the same size – as in the European Union – and an economic union of, say, France and Monaco, in which Monaco's economy is orders of magnitude smaller.

To take this difference into account, it makes sense to divide the countries into finitely many groups by size, so that the above the-same-prediction requirement be applicable only when by changing the values, we keep each country within the same group.

These groups should be reasonable from the topological viewpoint – e.g., we should require that each of the corresponding domains D of possible values is contained in a closure of its interior: $D \subseteq \overline{\text{Int}\,(D)}$, i.e., that each point on its boundary is a limit of some interior points.

Each domain should be strongly connected – in the sense that each two points in each interior should be connected by a curve which lies fully inside this interior.

Let us describe the resulting modified assumption in precise terms.

A precise description of the modified assumption. We assume that the set of all possible values of the input

$$v = \left(q_{1,t}, \ldots, e_{n,t-\ell_0}\right)$$

to the function f_i is divided into a finite number of non-empty non-intersecting strongly connected domains $D^{(1)}, \ldots, D^{(S)}$. We require that each of these domains is contained in a closure of its interior $D^{(s)} \subseteq \overline{\text{Int}\left(D^{(s)}\right)}$. We then require that if the following conditions are satisfied for the fours inputs $v^{(1)}$, $v^{(2)}$, $u^{(1)}$, and $u^{(2)}$:

- the inputs $v^{(1)}$ and $u^{(1)}$ belong to the same domain,
- the inputs $v^{(2)}$ and $u^{(2)}$ also belong to the same domain (which may be different from the domain containing $v^{(1)}$ and $u^{(1)}$), and
- we have $v^{(1)} + v^{(2)} = u^{(1)} + u^{(2)}$,

then we should have

$$f_i\left(v^{(1)}\right) + f_i\left(v^{(2)}\right) = f_i\left(u^{(1)}\right) + f_i\left(u^{(1)}\right).$$

Our main result. Our main result – proven in the next section – is that under the above assumption, the function $f_i(v)$ is piece-wise linear.

Discussion. This result explains why piece-wise linear models are indeed ubiquitous in econometrics.

Comment. Since the functions f_i are continuous, on the border between two zones with different linear expressions E and E', these two linear expressions should attain the same value. Thus, the border between two zones can be described by the equation $E = E'$, i.e., equivalently, $E - E' = 0$. Since both expressions are linear, the equation $E - E' = 0$ is also linear, and thus, describes a (hyper-)plane in the space of all possible inputs.

So, the zones are separated by hyper-planes.

3 Proof of the Main Result

1°. We want to prove that the function f_i is linear on each domain $D^{(s)}$. To prove this, let us first prove that this function is linear in the vicinity of each point $v^{(0)}$ from the interior of the domain $D^{(s)}$.

1.1°. Indeed, by definition of the interior, it means that there exists a neighborhood of the point $v^{(0)}$ that fully belongs to the domain $D^{(s)}$. To be more precise, there exists an $\varepsilon > 0$ such that if $|d_q| \le \varepsilon$ for all components d_q of the vector d, then the vector $v^{(0)} + d$ also belongs to the domain $D^{(s)}$.

Thus, because of our assumption, if for two vectors d and d', we have

$$|d_q| \le \varepsilon, \ \ |d'_q| \le \Delta, \text{ and } |d_q + d'_q| \le \varepsilon \text{ for all } q, \tag{4}$$

then we have

$$f_i\left(v^{(0)} + d\right) + f_i\left(v^{(0)} + d'\right) = f_i\left(v^{(0)}\right) + f\left(v^{(0)} + d + d'\right). \tag{5}$$

Subtracting $2f_i\left(v^{(0)}\right)$ from both sides of the equality (5), we conclude that for the auxiliary function

$$F(v) \stackrel{\text{def}}{=} f_i\left(v^{(0)} + v\right) - f_i\left(v^{(0)}\right), \tag{6}$$

we have

$$F\left(d + d'\right) = F(d) + F\left(d'\right), \tag{6}$$

as long as the inequalities (4) are satisfied.

1.2°. Each vector $d = (d_1, d_2, \ldots)$ can be represented as

$$d = (d_1, 0, \ldots) + (0, d_2, 0, \ldots) + \ldots \tag{7}$$

If $|d_q| \le \varepsilon$ for all q, then the same inequalities are satisfied for all the terms in the right-hand side of the formula (7). Thus, due to the property (6), we have

$$F(d) = F_1(d_1) + F_2(d_2) + \ldots, \tag{8}$$

where we denoted

$$F_1(d_1) \stackrel{\text{def}}{=} F(d_1, 0, \ldots), \;\; F_2(d_2) \stackrel{\text{def}}{=} F(0, d_2, 0, \ldots), \ldots \tag{9}$$

1.3°. For each of the functions $F_q(d_q)$, the formula (6) implies that

$$F_q\left(d_q + d_q'\right) = F_q(d_q) + F_q\left(d_q'\right). \tag{10}$$

In particular, when $d_q = d_q' = 0$, we conclude that $F_q(0) = 2F_q(0)$, hence that

$$F_q(0) = 0.$$

Now, for $d_q' = -d_q$, formula (10) implies that

$$F_q(-d_q) = -F_q(d_q). \tag{11}$$

So, to find the values of $F_q(d_q)$ for all d_q for which $|d_q| \le \varepsilon$, it is sufficient to consider the positive values d_q.

1.4°. For every natural number N, formula (10) implies that

$$F_q\left(\frac{1}{N} \cdot \varepsilon\right) + \ldots + F_q\left(\frac{1}{N} \cdot \varepsilon\right) (N \text{ times}) = F_q(\varepsilon), \tag{12}$$

thus

$$F_q\left(\frac{1}{N} \cdot \varepsilon\right) = \frac{1}{N} \cdot F_q(\varepsilon). \tag{13}$$

Similarly, for every natural number M, we have

$$F_q\left(\frac{M}{N} \cdot \varepsilon\right) = F_q\left(\frac{1}{N} \cdot \varepsilon\right) + \ldots + F_q\left(\frac{1}{N} \cdot \varepsilon\right) (M \text{ times}),$$

thus

$$F_q\left(\frac{M}{N} \cdot \varepsilon\right) = M \cdot F_q\left(\frac{1}{N} \cdot \varepsilon\right) = M \cdot \frac{1}{N} \cdot F_q(\varepsilon) = \frac{M}{N} \cdot F_q(\varepsilon).$$

So, for every rational number $r = \dfrac{M}{N} \le 1$, we have

$$F_q(r \cdot \varepsilon) = r \cdot F_q(\varepsilon). \tag{14}$$

Since the function f_i is continuous, the functions F and F_q are continuous too. Thus, we can conclude that the equality (14) holds for all real values $r \le 1$.

By using formula (11), we can conclude that the same formula holds for all real values r for which $|r| \le 1$.

Now, each d_q for which $|d_q| \le \varepsilon$ can be represented as $d_q = r \cdot \varepsilon$, where $r \stackrel{\text{def}}{=} \dfrac{d_q}{\varepsilon}$. Thus, formula (14) takes the form

$$F_q(d_q) = \frac{d_q}{\varepsilon} \cdot F_q(\varepsilon),$$

i.e., the form

$$F_q(d_q) = a_q \cdot d_q, \tag{15}$$

where we denoted $a_q \stackrel{\text{def}}{=} \dfrac{F_q(\varepsilon)}{\varepsilon}$. Formula (8) now implies that

$$F(d) = a_1 \cdot d_1 + a_2 \cdot d_2 + \ldots \tag{16}$$

By definition (6) of the auxiliary function $F(v)$, we have

$$f_i\left(v^{(0)} + d\right) = f_i\left(v^{(0)}\right) + F(d),$$

so for any v, if we take $d \stackrel{\text{def}}{=} v - v^{(0)}$, we would get

$$f_i(v) = f_i\left(v^{(0)}\right) + F\left(v - v^{(0)}\right). \tag{17}$$

The first term is a constant, the second term, due to (16), is a linear function of v, so indeed the function $f_i(v)$ is linear in the ε-vicinity of the given point $v^{(0)}$.

2°. To complete the proof, we need to prove that the function $f_i(v)$ is linear on the whole domain. Indeed, since the domain $D^{(s)}$ is strongly connected, any two points are connected by a finite chain of intersecting open neighborhood.

In each neighborhood, the function $f_i(v)$ is linear, and when two linear function coincide in the whole open region, their coefficients are the same. Thus, by following the chain, we can conclude that the coefficients that describe $f_i(v)$ as a locally linear function are the same for all points in the interior of the domain.

Our result is thus proven.

Acknowledgments

This work was supported by Chiang Mai University, Thailand. We also acknowledge the partial support of the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand, and of the US National Science Foundation via grant HRD-1242122 (Cyber-ShARE Center of Excellence).

The authors are greatly thankful to Professor Hung T. Nguyen for his help and encouragement.

References

1. J. Aczél and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, 2008.
2. T. Bollerslev, R. Y. Chou, and K. F. Kroner, "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence", *Journal of Econometrics*, 1992, Vol. 52, pp. 5–59.
3. P. J. Brockwell and R. A. Davis, *Time Series: Theories and Methods*, Springer Verlag, New York, 2009.
4. W. Enders, *Applied Econometric Time Series*, Wiley, New York, 2014.
5. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.
6. L. R. Glosten, R. Jagannathan, and D. E. Runkle, "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", *Journal of Finance*, 1993, Vol. 48, pp. 1779–1801.
7. H. T. Nguyen, V. Kreinovich, O. Kosheleva, and S. Sriboonchitta, "Why ARMAX-GARCH Linear Models Successfully Describe Complex Nonlinear Phenomena: A Possible Explanation", In: V.-N. Huynh, M. Inuiguchi, and T. Denoeux (eds.), *Integrated Uncertainty in Knowledge Modeling and Decision Making, Proceedings of The Fourth International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making IUKM'2015*, Nha Trang, Vietnam, October 15–17, 2015, Springer Lecture Notes in Artificial Intelligence, 2015, Vol. 9376, pp. 138–150.
8. R. S. Tsay, *Analysis of Financial Time Series*, Wiley, New York, 2010
9. J. M. Zakoian, *Threshold heteroskedastic models*, Technical Report, Institut National de la Statistique et des Études Économiques (INSEE, 1991.
10. J. M. Zakoian, "Threshold heteroskedastic functions", *Journal of Economic Dynamics and Control*, 1994, Vol. 18, pp. 931–955.