Journal of Econometrics 167 (2012) 274–294

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

# Likelihood estimation and inference in threshold regression

Ping Yu

*Department of Economics, 12 Grafton Road, University of Auckland, New Zealand*

ARTICLE INFO

*Article history:*
Received 18 January 2009
Received in revised form
10 April 2011
Accepted 14 December 2011
Available online 21 December 2011

*JEL classification:*
C11
C13
C21

*Keywords:*
Threshold regression
Structural change
Nonregular models
Boundary
Efficiency bounds
Bayes
Middle-point MLE
Compound Poisson process
Wiener–Hopf equation
Local asymptotic minimax
Credible set

ABSTRACT

This paper studies likelihood-based estimation and inference in parametric discontinuous threshold regression models with i.i.d. data. The setup allows heteroskedasticity and threshold effects in both mean and variance. By interpreting the threshold point as a "middle" boundary of the threshold variable, we find that the Bayes estimator is asymptotically efficient among all estimators in the locally asymptotically minimax sense. In particular, the Bayes estimator of the threshold point is asymptotically strictly more efficient than the left-endpoint maximum likelihood estimator and the newly proposed middle-point maximum likelihood estimator. Algorithms are developed to calculate asymptotic distributions and risk for the estimators of the threshold point. The posterior interval is proved to be an asymptotically valid confidence interval and is attractive in both length and coverage in finite samples.



## 1. Introduction

Since its invention by Tong Howell in the 1970s, the threshold regression model is popular in both statistics and econometrics.[1] Particularly, it has many applications in economics, e.g., Potter (1995), Durlauf and Johnson (1995), Savvides and Stengos (2000), Huang and Yang (2006) and Boetel et al. (2007) among others; see also Lee and Seo (2008) for other examples. The typical setup of threshold regression models is

$$y = \begin{cases} x'\beta_1 + \sigma_1 e, & q \le \gamma; \\ x'\beta_2 + \sigma_2 e, & q > \gamma; \end{cases} \tag{1}$$
$$E[e|x, q] = 0,$$

where $q$ is the threshold variable used to split the sample, $\gamma$ is the threshold point, $x \in \mathbb{R}^k$, $\beta \equiv (\beta_1', \beta_2')' \in \mathbb{R}^{2k}$ and $\sigma \equiv (\sigma_1, \sigma_2)'$ are threshold parameters on the mean and variance in the two regimes. We set $E[e^2] = 1$ as a normalization of the error variance and allow for conditional heteroskedasticity. All the other variables have the same definitions as in the linear regression framework.

There are two asymptotic frameworks for statistical inference on $\gamma$. The first is introduced by Chan (1993) in a nonlinear time series context, where $(\beta_1', \sigma_1)' - (\beta_2', \sigma_2)'$ is a fixed constant. The second is introduced by Hansen (2000), where no threshold effect on variance exists and the threshold effect in mean diminishes asymptotically. This paper uses the discontinuous framework of Chan (1993) with i.i.d. data. The results developed in this paper can serve as a benchmark for more complicated data generating processes in time series and panel data.

Both Chan (1993) and Hansen (2000) use least squares criteria to estimate $\gamma$, and derive the asymptotic distributions of the corresponding least squares estimators (LSEs), but the efficiency theory has never been studied. As Andrews (1993) concludes in the related structural change context, "no optimality properties are known for the ML estimator of $\pi$", where $\pi$ is the structural change point and plays a similar role as $\gamma$ in threshold regression. This paper intends to fill this gap in a parametric setting. In this environment, the density of $e$ conditional on $(x, q)$ is assumed to be $f_{e|x,q}(e|x, q; \alpha)$, where $\alpha \in \mathbb{R}^{d_\alpha}$ is some nuisance parameter affecting the shape of the error distribution. The joint distribution

*E-mail addresses:* p.yu@auckland.ac.nz, whistle.yu@gmail.com.

[1] See Howell (2007) for the birth of the threshold time series model.

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.12.002

of $(x, q)$ is $f_{x,q}(x, q)$, the marginal distribution of $q$ is $f_q(q)$, and the unknown parameter is $\theta = \left(\beta_1', \beta_2', \sigma_1, \sigma_2, \alpha', \gamma\right)' \equiv \left(\underline{\theta}', \gamma\right)'$.

In regular models, it is well known that the Bayes estimator (BE) and the maximum likelihood estimator (MLE) are asymptotically equivalent; see, e.g., Theorem 10.8 of Van der Vaart (1998). In nonregular models, however, Hirano and Porter (2003) and Chernozhukov and Hong (2004) show that the BE can be more efficient than the MLE in boundary estimation. By interpreting $\gamma$ as a "middle" boundary of $q$, this paper finds a similar result about the estimation of $\gamma$. It is worth pointing out that the threshold regression model is more general than the models in the above-mentioned boundary literature. As illustrated in the following Section 2, the conventional boundary problems are special cases of (1) in an extremely simplified setup. Like the usual boundary literature, the results of this paper are developed using the framework in the seminal book by Ibragimov and Has'minskii (1981). Their Chapter 5 also discusses the statistical inference when densities have jumps, but their arguments seem more relevant to Hirano and Porter (2003) and Chernozhukov and Hong (2004).

The models considered in this paper are very general. For example, we allow heteroskedasticity and threshold effects in both mean and variance, error distributions with general parametric forms, and general loss functions. Independent work by Chan and Kutoyants (2010) considers a problem similar to this paper in threshold autoregressive models under very restrictive specifications; e.g., the error term is i.i.d. normal, the slope parameters are known, and only the mean square error loss is considered. They use a simulation method to find critical values for the confidence intervals (CIs) of the threshold point, which is, as argued in Section 3.4, not practical in reality. Instead, we suggest to use the posterior interval as the CI for the threshold point. Most importantly, they give little details on the efficiency problem.

This paper is organized as follows. Section 2 illustrates the main idea of this paper using a simple threshold regression model. Section 3 presents the main result of this paper, in which the asymptotic distributions of the MLE and BE are derived, and the BE is proved to be most efficient among all estimators. Also, the posterior interval is proven to be an asymptotically valid confidence interval. Section 4 shows some simulation results, and Section 5 concludes. All assumptions, proofs, lemmas and algorithms are given in Appendices A–D, respectively.

Before closing this introduction, it should be pointed out that the framework of this paper is essentially frequentist in the sense that while Bayes procedures are used, the randomness is confined to the data and does not include parameters. Correspondingly, we do not intend to propose a new Bayesian simulation method; such methods can be found in Geweke and Terui (1993). A word on notation: the letter $C$ is used as a generic positive constant, which need not be the same in each occurrence. $\ell$ is always used for indicating the two regimes in (1), so is not written out explicitly as "$\ell = 1, 2$" throughout the paper. The code for simulations is available at http://homes.eco.auckland.ac.nz/pyu013/research.html.

## 2. No error term: an illustration

In this section, a simple threshold regression model is used to illustrate the main result of this paper: the threshold point is essentially a "middle" boundary. In the following discussion, $q_{(m)}$ denotes the $m$th order statistic of a sequence of random variables $\{q_i\}_{i=1}^n$.

Suppose the population model is

$$y = \mathbf{1}(q \le \gamma), \quad q \sim U[0, 1], \tag{2}$$

where $U[0, 1]$ is the uniform distribution on $[0, 1]$, $\mathbf{1}(\cdot)$ is the indicator function, $\gamma$ is the parameter of interest, and $\gamma_0 = 1/2$.[2] This is equivalent to $x = 1$, $\beta_{10} = 1$, $\beta_{20} = 0$, and $\sigma_{10} = \sigma_{20} = 0$ in the general setup (1). There is no error term $e$ in (2), so the observed $y$ value can only be 0 or 1. Such a simple model can be viewed as a treatment rule in social program evaluation. If $q$ is interpreted as the percentiles of income, then people below the median income are enrolled in the program with $y$ taking value 1. Otherwise, people are not enrolled with $y$ taking value 0. Such a treatment rule is too simple in reality, as the propensity score is a step function dropping from 1 ($q$ below $\gamma_0$) to 0 ($q$ above $\gamma_0$).[3] Here, the task is to find the step treatment rule, given the income of people and whether they are enrolled in the program.

For this simple model, the likelihood function is

$$p\left(W_n|\gamma\right) = \prod_{i=1}^n \left(\mathbf{1}\left(q_i \le \gamma\right)^{\mathbf{1}(y_i=1)} \cdot \mathbf{1}\left(q_i > \gamma\right)^{\mathbf{1}(y_i=0)}\right), \tag{3}$$

where $W_n = (w_1, \ldots, w_n)$ with $w_i \equiv (y_i, q_i)$ is the dataset, and $0^0$ is defined to be 1. A simple calculation shows that the MLE is $[q_{(m)}, q_{(m+1)})$, where $m$ is the number of $y_i$'s with value 1. When there is an interval maximizing this likelihood function, following the literature, the left endpoint (i.e., $q_{(m)}$) is taken as the estimator.[4] Such an estimator is called the left-endpoint MLE (LMLE), and is denoted as $\widehat{\gamma}_{\text{LMLE}}$.

First, $\widehat{\gamma}_{\text{LMLE}}$ is $n$ consistent. Notice that $\widehat{\gamma}_{\text{LMLE}}$ is the $q_i$ that is closest to $1/2$ from the left. Since $\{q_i\}_{i=1}^n$ are sampled from $U[0, 1]$, $n$ data points of $q$ are randomly put into an interval with length 1, and thus the average distance between contiguous $q_i$'s is around $1/n$. $1/2$ is in the interval $[q_{(m)}, q_{(m+1)})$, so $n(\widehat{\gamma}_{\text{LMLE}} - 1/2)$ is expected to be $O_p(1)$. Second,

$$n(\widehat{\gamma}_{\text{LMLE}} - 1/2) \xrightarrow{d} -\text{Exp}(1), \tag{4}$$

where Exp(1) is a standard exponential distribution. Since $\widehat{\gamma}_{\text{LMLE}}$ is smaller than $1/2$, for any $t \le 0$,

$$\begin{aligned} P\left(n(\widehat{\gamma}_{\text{LMLE}} - 1/2) \le t\right) &= P\left(q_i \notin \left(1/2 + \frac{t}{n}, 1/2\right] \text{ for all } i\right) \\ &= \left(1 + \frac{t}{n}\right)^n \to e^t. \end{aligned}$$

To further appreciate (4), suppose we want to estimate $\gamma$ in the distribution of $yq$. $yq$ picks out the $q$'s such that $y = 1$. Its distribution is a point mass at 0 plus a uniform distribution on $(0, \gamma]$. Since $\gamma$ is the right endpoint of this distribution, it is well known that the MLE is the maximum of the data and follows the exponential distribution asymptotically. Similarly, $\gamma$ can be treated as the left boundary of $(1 - y)q$ and estimated by the minimum of the nonzero $(1 - y_i)q_i$. In short, $\gamma$ can be viewed as a boundary (of both $yq$ and $(1 - y)q$) although it is in the middle of $q$'s support.

2 I would like to thank Jack Porter for providing this example. I also want to thank an associate editor and a referee for improving its exposition.

3 Such a treatment rule is called the sharp design, as opposed to the fuzzy design where the treatment is not deterministic in the two regimes, by Trochim (1984) in the regression discontinuity design (RDD) literature; see Bajari et al. (2010) for a similar analysis as below when using RDD to study contracting in health care. Usually, $\gamma_0$ is set by the policy-maker, and is publicly known.

4 Most of the literature uses the left endpoint instead of the middle point. A possible reason is that these two estimators are thought to bear similar properties. For example, the sample splitting based on either point is the same; the maximizing interval shrinks at rate $1/n$ as shown in the following paragraph, so both methods generate almost the same point estimate in practice. The only exception to use the middle point, to my knowledge, is Gijbels et al. (1999) in the nonparametric environment, but they do not provide any theoretical justification.