Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 28 (2025), No. 2, pp. 375–385
DOI : 10.47974/JSMS-1395

Estimating switch points and thresholds when the dependent variable is not observable

T.V.S. Ramamohan Rao
Department of Economics
Indian Institute of Technology
Kanpur 208016
Uttar Pradesh
India

Abstract

This study deals with a mixture of switching and threshold regression when the dependent variable cannot be measured. However, unlike the earlier attempts, the proposed methods do not assume an *apriori* probability density function for any of these variables. The switch points and thresholds have been estimated separately. Similarly, the estimation procedure defines transformation of independent variables to reduce the problem to one of estimating a continuous linear regression. Examples have been provided to illustrate the efficacy of the proposed method.

Subject Classification: *62P20,62J99, 91B24.*

Keywords: *Unobservable dependent variable, Switching and threshold, Efficient estimation.*

1. Introduction

"We can observe people's actions, but we cannot observe their motives. We don't know what people are thinking or what they are trying to achieve."- Heath [4, p.31].

A periodic review meeting of the board of directors of most firms takes stock of (a) its performance – measured by its sales, profit, finances attracted from the stock market and so on, (b) the compliance of the workers and management with the tasks and objectives defined for them, and (c) chart a course of action for the next time period. A variety of considerations are taken into account, several strategic choices (both quantitative and qualitative) are

ORCID-ID: `0000-0002-1632-2691`

E-mail: `rmrao@iitk.ac.in`

announced, and define a combination of goals to be achieved. Clearly, it is difficult to identify a single objective or a well defined combination of goals. From an analytical viewpoint it is difficult, if not impossible, to identify a single measurable objective that drives managerial decisions.

Management studies, including the emergence of machine learning algorithms, define a single measurable objective and explain its attainment through the use of finite and well defined strategic choices. It would be more appropriate to consider the dependent variable as non-measurable.

The external environment and the actions and reactions of rival firms are constantly changing. As a result the performance of a firm depends on the switches experienced by a given strategy of the firm. On occasions, a strategy may attain a threshold and further progress of the firm necessitates a new strategic choice. The process repeats cyclically.

Most analytical studies of these experiences postulate a single measurable objective. They also deal with switches or thresholds in the performance independently. They are never treated jointly with a threshold occurring after a switch and inducing another strategic choice that initiates a switch away from the threshold. This study offers a synthetic approach to such a phenomenon.

The switch points and thresholds are not known *apriori.* They need to be estimated. The existing studies, based on statistical inference, postulate that these changes are random variables with a predefined probability distribution. Similarly, the non-measurable objective is treated as a random variable. It can be readily inferred that the resulting conclusions of the studies reflect the properties of the probability distribution postulated. The present study suggests that it is not necessary to specify such randomness *apriori.*

A more direct approach to the identification of the points of discontinuity, defining transformation of variables to reduce the nonlinear regression problem to a single continuous linear regression will be advocated. Clearly, the number of patterns of discontinuity cannot be postulated initially. Instead, they must be estimated from the available data. The important contribution of this study it to make the determination of a specific pattern operational, applicable to the given data, and specify the process of estimation even when a non-measurable dependent variable should be taken into account.

2. Structure of Previous Analytical attempts[1]

Consider the simplest threshold regression problem

$$V = \text{as for } s \leq s^*$$
$$= \text{as}^* \text{ for } s > s^*$$

In much of the analysis the threshold s^* is postulated to be random. To simplify the presentation assume that it is uniformly distributed over (0, $2m$).

[1] Very few studies dealing with an unobserved dependent variable are available in the literature. An exhaustive search revealed the following prominent studies: Dell'Anmo [1], El-Dosokey [2], Gunny et al [3], Marra et al [5], Rao [6], and Yu [7]. It was not possible to identify any common method of specification or estimation. Hence, an exhaustive review is not attempted.

Let u be a random realization of s^*. Then, $s \leq s^*$ is equivalent to $u \geq s$. The probability of such an event is

$$p \leq s^* = prob\ (u \geq s) = 1 - s/2m$$

Similarly,

$$p\ (s > s^*) = s/2m$$

Therefore, the expected value of V is

$$E(V) = as - bs^{2,}, \qquad \text{where } b = a/4m$$

The parameters a and b can be estimated by minimizing

$$Q = \sum\{V - (as - bs^2)\}^2$$

It results in

$$\sum Vs = a \sum s^3\ , \qquad \text{and}$$

$$\sum Vs^2 = a \sum s^3 - b \sum s^4$$

The estimate of m will then be $m = a/4b$

Note that the values of a and m depend on the probability distribution that is postulated. Secondly, note that the use of the quadratic approximation implicit in using $E(V)$ will not provide a satisfactory fit for the non-linearity implied by the threshold regression.

Consider the case where V is not observable. Usually the analysis proceeds by assuming that V is a random variable with a predefined probability distribution. Assume, for simplicity of exposition, that V is uniformly distributed over (0, 2m). Let u be a realization of the random variable.

The model can then be written as

$$u = \text{as for } s \leq s^*$$

$$= a\, s^* \text{ for } s > s^*$$

Note that given the assumption about the probability density function

$$E(u) = m, \text{and } \sigma^2\ (u) = m^2/3$$

Assume further that $prob\ (\ s \leq s^*) = p.$ Then,

$$E(U) = m = asp + a\ (1 - p)s^*$$

Consequently,

$$m - as^* = ap\ (s - s^*)$$

Hence,

(1) $ap = \sum(m - as^*)(s - s^*)/\sum(s - s^*)^2$

From the expected value it can be inferred that

(2) $mn = ap \sum s + a(1 - p)s^* n\text{m}$

with ap given by (1)

Denote by $A = ap$ and $B = a(q - p).$ It follows that

(3) $a = A + B,$ and $p = A/(A + B)$

Since

$$u - m = as - m \; for \; s \; \leq \; s^*$$
$$= as^* - \; m \; for \; s > \; s^*$$

(4) $\sigma^2 = m^2/3 \;=\; (as - m)^2 \, p + \; (as^* \; - m)^2 \; (1 - p)$

These four equations can be solved iteratively to estimate $a, p, s^*,$ $and \; m.$

Note, however, that the estimated values depend on the assumption about the distribution of the unobserved V. Further, only a linear approximation of the non-linear regression is indicated since the expected value assumption is crucial in the estimation process. Clearly, this will be an inefficient explanation of V.

The estimation method proposed in the following sections does not make any assumption about the randomness of s^* or V. A more direct method of identifying the switch points and thresholds will be outlined. Further, using simple transformations of the independent variables the non-linear problem will be converted into a continuous linear regression. For all practical purposes the proposed generalization is a fundamental approach to the problem at hand.

3. Reconstructing the Model

The restructuring of the problem will be defined for one dependent variable that is not measurable and one explanatory variable that experiences one switch and a threshold in the sample. It is easy to visualize a generalization to multiple switch points and many thresholds.

Let L denote the extent of consumer loyalty to the products of a firm depending on the (a) perception of the consumers about the quality and usefulness of the product relative to those of rival firms on the market, and (b) the effectiveness of the advertising campaigns (A) of the firm. It can be surmised that L will have a threshold based on these considerations. In practice, it can be noted that consumers tend to ignore advertisements of the firm once they recognize the value of the product. Thus, both L and A attain their respective thresholds. This is depicted in Fig. 1.

In general, the choice of A by the firm depends on the cost of doing so relative to the revenue that the firm can generate. Let

$$C = cost = \; c_0 + \; c_1 \, A + \; c_2 S, \text{ where}$$
$$S = \text{volume of sales}$$

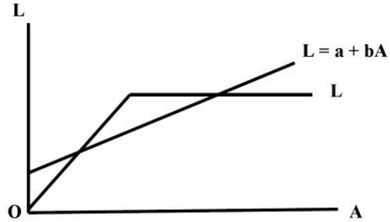

Figure 1
Thresholds on A and L

Note that c_2 is generally positive. Hence, the choice of A will be lower as S increases. Two observations are pertinent: if S is low, and consequently the cost, the firm would prefer $A > A^*$ (the threshold value). The expected loyalty would, conceptually, exceed the cost of using A. However, the costs will go up with the increase in S and A will be reduced.

Consider the threshold regression including a switch

$$L = aA \; for \; A \le A_1^*$$
$$= bA \; for \; A_1^* < A \le A_2^*$$
$$= bA_1^* \; for \; A > A_2^*$$
$$C = c_0 + c_1 A + c_2 S$$

For A $\le A_1^*$ the choice of A would be such that

$$aA = c_0 + c_1 A + c_2 S$$

That is,

$$A = (c_0 + c_2 S)/(a - c_1)$$

Similarly,

$$A = (c_0 + c_2 S)/(b - c_1) \; for \; A_1^* < A \le A_2^*$$

When $A > A_2^*$ the choice of S would depend on A_2^* and the corresponding S^* . Note that even if the S chosen is greater than S^* the resulting L would only depend on S^* and A_2^*.

Consequently, the choice of A^* would satisfy the equation

$$C^* = bA_2^* = c_0 + c_1 A_2^* + c_2 S^*$$

That is, the resulting A_2^* would be

$$A_2^* = (c_0 + c_2 S^*)/(b - c_1)$$

Define $y = (c_0 + c_2 S)$. The, the resulting threshold regression can be written as

$$A = y/(a - c_1) \text{ for } y \le y_1$$
$$= y/(b - c_1) \text{ for } y_1 < y \le y_2$$
$$= y^*/(b - c_1) \text{ for } y > y_2$$

Note that the cost function and the associated parameters can be estimated independently.

Reconsider the above problem. Note that when S is high, the low value of A is such that

$$C = c_0 + c_1 A + c_2 \mathrm{S} = aA$$

That is,

$$S = \{ -c_0 + (a - c_1) A \} / c_2$$

Similarly, A^* will satisfy the equation

$$S^* = \{ -c_0 + (a - c_1) A^* \} / c_2$$

The model can be equivalently written as

$$S + (c_0/\, c_2) = (a - c_1)A \text{ for } A \le A_1^*$$
$$= (b - c_1)A \text{ for } A_1^* < A \le A_2^*$$
$$= (b - c_1)\, A_2^* \text{ for } A > A_2^*$$

The following observations are pertinent. First, $C = L$ will be satisfied for every choice of A and S. Second, the threshold values and switch points of $A, S, C,$ and L will be the same corresponding to A^*. Accordingly, the model can also be written as

$$C = aA \text{ for } A \le A_1^*$$
$$= bA \text{ for } A_1^* < A \le A_2^*$$
$$= bA_2^* \text{ for } A > A_2^*$$

Some extensions can also be conceptualized. First, the cost will depend on the level of inventories as well. Similarly, the extent of debt (borrowing) will have an effect on cost. The model can be extended to account for all such variables as well. Second, the management may have several objectives in perspective though the exact choice and the weights assigned to different objectives cannot be stated *aprior*i. Third, for all practical purposes, C is used as a proxy for L. It would be prudent to accept the estimates based on C until a better alternative can be conceptualized given that L is not measurable.

The difficulty is in finding out (a) what motivates mangers to make changes in the nature (qualitative) and quantity of their strategies, (b) how they conceptualize changes, and (c) the mechanisms through which they implement their strategies. The information available (big data) is perforce limited and no amount of quantitative data that can be assembled, statistical averaging, and machine learning can disentangle such a complex phenomenon.

4. Methods of Estimation

In much of the existing literature the attempt has been to estimate $a, s,$ and A^* simultaneously. However, it would be more pragmatic to assume that the management initially visualizes the emergence of a switch or a threshold and designs strategies accordingly. In general, the management does not acknowledge that a threshold is imminent when the performance is improving. They tend to switch strategies only when it becomes difficult to achieve anything near the expected performance. The switches are invariably new strategies. However, the available data cannot identify the new strategy. Similarly, statistical inference may not be in a position to specify the length of time over which a strategy remains efficient. The statistical methods, that have been presented in the literature as well as the methods detailed below, may be inadequate to the task for all their sophistication.

The primary task of the management is to identify the switch points and thresholds as they emerge. The potential of the strategies to explain performance will follow once the change points are clearly demarcated. This task can be approached in the following manner. Assume, to begin with, that

$$S = A_0 + a_1 A + error$$

without any discontinuity. The error of such estimation will be

$$E = S - SE \text{ , where}$$

$$SE = A_0 + a_1 A$$

It will exhibit the patterns illustrated in Fig. 2 depending on whether a point of change is a switch or a threshold. Since the available data is not continuous the points of change may be marked by sudden shifts.

The management may be utilizing a combination of strategies to achieve their goal. In such cases the combination of strategies that explain the discontinuities may be represented by the principal components of such strategies. Two or more components may explain the switch points or thresholds. The principal components are orthogonal to each other and enable the identification of the changes in different components without referring to the others[2].

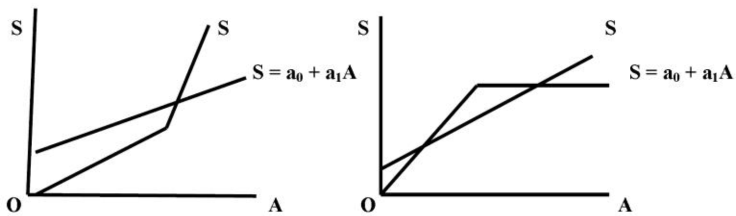

Figure 2
Estimating A_s and A_t

One other consideration needs attention. The management may consider a change in strategy when the current choice reaches a particular value irrespective of the time at which a change is necessitated. On the other hand, whatever the past experience may have been, the management takes the current position into account and chooses a proper strategy. The later approach is more pragmatic. It will be the choice in the following analysis.

Assume that the switch points and thresholds of every strategy (or an appropriate combination) have been identified[3]. The estimation procedure

[2] It can be argued that in the final analysis the identification of switch points and thresholds depends on the specific method adopted. Hence, it cannot be claimed that the method proposed here is superior to assuming that the changes are random with a predefined probability density function. Statistical methods of inference, irrespective of the form they take, based on an analysis of *ex* post data may not reflect the actual choices of the management.

[3] Trivially, it can be argued that there is switch at every observation so that $y_i = a_i x_i$ for all i, where $a_i = y_i / x_i$ provides the best fit. (However, threshold regressions can never be that trivial). This invalidates the spirit of scientific generalization. Such a concept is based on the notion that there is consistency of behavior of an individual over time. Similar is the argument that there is uniformity of behavior among various individuals. Further, it may not be possible to capture all the non-linear effects in the regression if only a few points of change can be identified. In general, the optimum number of switch points necessarily depends on the implied decision process. It is necessary to initially disentangle it. As a result, the specification of the optimum number of points of change remains open. Statistical inference can be helpful only after that process is completed.

detailed below is quite general though it will be illustrated with one switch and one threshold in the model.

Consider the model defined earlier. Construct the following variables.

$$A_1 = A \text{ for } A \le A_1^*$$
$$= 0 \text{ thereafter}$$
$$A_2 = 0 \text{ for } A_1^* \le A \le A_2^*$$
$$= A_2^* \text{ for } A > A_2^*$$

Estimate the model

$$S = a\,A_1 + bA_2 + error$$

as a continuous regression. This method accommodates the non-linearity alluded to above.

5. Illustrative Examples

This section provides two examples to illustrate the working of the proposed method of estimation. The first is constructed to check the robustness of the method. The second is based on the data for a firm in the textile industry in India. The data was obtained from the Center for Monitoring the Indian Economy (CMIE).

The data for example 1 is provided in Table 1.

Table 1
Sample Data

S	A	C	Error	
1.092	.105	.81	.2543	
1.248	.12	.91	.1052	
1.300	.125	.94	.0555	
1.610	.155	.97	-.2447	
1.650	.173	1.01	-.5708	
1.710	.185	1.22	-.6328	
1.770	.197	1.25	-.6835	
2.840	.181	1.82	.4158	switch
2.920	.185	1.86	.4551	
3.080	.189	1.96	.5338	threshold
3.09	.210	1.97	.1166	
3.110	.212	1.99	.0959	
3.12	.214	2.12	.0653	
3.15	.217	2.23	.0342	

The estimated cost function was

$$c_0 = .0280,\ c_1 = .7995, \text{ and } c_2 = .5917$$

Estimating $C = f(A)$ yielded the following values for the parameters

$$a_1 = 6.7673,\ a_2 = 10.6293, \text{ and } GFit = .9957$$

The same parameters when estimated using $S = f(A)$ provided the following results

$$A_1 = 158, \text{ and } A_2 = .1109$$

$$a_1 = 6.8305, a_2 = 9.8124, \text{ and } GFit = .9976$$

The data for Arvee Denim Ltd is provided in Table 2.

Table 2
Data for Arvee Denim

S	A	C	Error	
971.1	3.0	975.4	69.1943	
1067.2	18.8	1039.2	-461.1242	
1495.2	37.4	1411.7	-770.5536	
1480.1	50.9	1226.2	-1320.8850	Switch
1729.3	43.6	1500.4	-782.2634	
2373.5	81.3	2003.3	-1631.746	Switch
2774.4	55.6	2276.3	-209.9243	
2499.1	32.3	2010.9	435.5448	Switch
2459.7	42.1	2199.7	7.6067	
3281.1	80.9	3040.1	-709.2869	Switch
3795.3	56.5	3302.8	772.2936	
4952.9	95.4	4372.6	387.6353	
6623.1	109.4	6172.7	1502.7809	
7236.8	102.9	6483.7	2374.1848	Threshold
6856.4	121.8	6481.3	1244.4613	
7324.4	141.5	6990.6	931.4205	
8413.5	240.8	8019.2	-1916.4010	Switch
8441.1	191.3	8063.2	760.7127	

To begin with, note that the errors were estimated after running the regression $S = f(A)$ ignoring the non-linearity. The result was

$$S = 782.9855 + 39.6468\ A,\ GFit = .9498$$

The estimated cost function was

$$C = -169.4379 + 2.1705\ A + .9128\ S,\ GFit = .9822$$

The following estimates were obtained when $S = f(A)$ was estimated using the cost function approach:

a = 44.7977, b = 33.5580, c = 35.8178, d = 65.4453, e = 61.0889, and f = 6.9100, $GFit$ = .9811

When the parameters were estimated directly using the cost function the following results emerged.

a = 42.7325, b = 28.4589, c = 29.8099, d = 55.9575, e = 55.6747, f = 8.8991, $GFit$ = .9822

Clearly, both the approaches indicate that the methods of estimation are viable.

6. Retrospect

The primary achievement of this study was in outlining a method of estimating the parameters of a regression comprising of switch points and threshold levels of the dependent variable even when it is not observable. It was postulated that the chosen levels of advertising and sales will be contingent on the cost of using such strategies. The proposed method was also efficient in isolating the switch points and thresholds. It was demonstrated both theoretically and empirically that such an approach to the unobservable dependent variable has the prospect of estimating the desired parameters satisfactorily.

However, several issues require further analysis. First, the loyalty of consumers to the products of a firm generally depends not only on its advertising campaigns but also on several other factors such as the level of inventories and the promptness of delivery of orders being a few of them. The proposed method can also take these into account if each of the causative factors has independent effects. Some experimentation to ensure minimization of the error sum of squares in estimation may be necessary. Second, the loyalty of consumers can itself be multidimensional. The following method may be useful. Let $y_1, y_2, \dots, y_n$ be some of the measurable alternatives. Define $T_1 = y_1$ and $T_m = y_m + a_m T_{m-1}$ for m = 2,3, ..., N where $a_m^2 = 1 - r_m^2\ (y_m, T_{m-1})$ and r_m is the correlation coefficient between y_m and T_{m-1}. How an optimum choice can be identified remains an open question. Third, in practice, the consumer loyalty, demand for the products of the firm, and the profits generated etc., depend not only on the strategies of the firm but also on external factors such as the market, rival firms reactions and so on. In any case, some values of the observed data will remain outliers to the proposed regression model. They cannot be explained by calculated management decisions based on logic. On the whole, the analytical process will be circumscribed by the logic of the decision processes and the availability of big data that necessarily includes several qualitative aspects. At the present stage it can only be concluded that a useful beginning to approach the problem on hand is available. Refinements will necessarily be an ongoing process.

References

[1] R. Dell' Anmo, "Measuring the unobservable: estimating informal economy by structural equation modeling approach," *International Tax and Public Finance*, vol. 30, pp. 247-277 (2023).

[2] A. El-Desokey, "Influence of unobserved data in the time series of the dependent variable and their position of analysis of multiple linear regression on prediction – case study on factors affecting CO2 emissions," *Advances in Mathematics: Scientific Journal*, vol. 12, pp. 329-344 (2023).

[3] Y. Guney, Y. Tual, S. Ozdemir, G. Yavuz, and O. Arslan, “An analysis to identify the structural breaks of Covid-19 in Turkey,” *Journal of Statistics and Management Systems*, vol. 25, pp. 1971-1989 (2022).

[4] J. Heath, *Economics without illusions*, New York: Broadway Books (2010).

[5] G. Marra, R. Radice, and D. Zimmer, “A unifying switching regime analysis with applications in health economics,” *Economics Reviews*, vol. 43, pp. 52-70 (2024).

[6] T. V. S. R. Rao, “On the estimation of a class of threshold regression models,” *Journal of Quantitative Economics*, vol. 22, pp. 199-209 (2024).

[7] K. Yu, “A linear regression with an unobserved dependent variable,” *Communications in Statistics: Theory and Methods*, vol. 17, pp. 3075-3087 (1988).

Received April, 2024