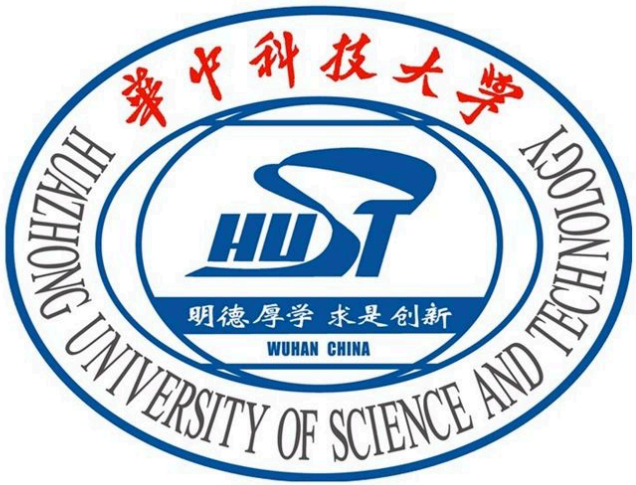
Threshold Regression Model

Zichao Yang

School of Economics, Huazhong University of Science and Technology

Date: May 27, 2016



What is Threshold Regression Model?

- Threshold regression models are used when we find that individual observations can be divided into classes based on the values of an observed variable.
- structural mutation problem exists widely in Economics
  - human capital and economic growth — S curve
  - government size and economic growth — Inverted U curve

How to cope with Threshold Regression Model?

- add square term of explanatory variable in your regression equation: $lm(y = x + x^2)$
- add dummy and cross term in your regression equation: $lm(y = x + dum + dum * x)$
- In his 1999 paper, *Threshold effects in non-dynamic panels: Estimation, testing, and inference*, Hansen introduced econometric techniques appropriate for threshold regression with panel data. Least squares estimation methods were described. An asymptotic distribution theory was derived which was used to construct confidence intervals for the parameters. A bootstrap method to assess the statistical significance of the threshold effect was also described.
- We will cover most of the above content in this note.

One Simple Model

- Suppose we know a very simple model as follow:

$$y_{i,t} = \begin{cases} \mu_i + x_{i,t}\beta_1' + \varepsilon_{i,t} & q_{i,t} \leq \gamma \\ \mu_i + x_{i,t}\beta_2' + \varepsilon_{i,t} & q_{i,t} > \gamma \end{cases} \tag{1}$$

  In this model, the observed data are from a balanced panel $\{y_{i,t}, q_{i,t}, x_{i,t} : 1 \leq i \leq n, 1 \leq t \leq T\}$. The subscript i indexes the individual and the subscript t indexes time. The dependent variable $y_{i,t}$ is scalar, the threshold variable $q_{i,t}$ is scalar, and the regressor $x_{i,t}$ is a k vector. Independent variable $x_{i,t}$ and threshold variable $q_{i,t}$ change over time and $\varepsilon_{i,t} \sim i.i.dN(0, \sigma_\epsilon^2)$. This model does not include lagged dependent variable as independent variable, hence, this is a non-dynamic model.
- Suppose $\beta = (\beta_1', \beta_2')'$, then the equation can be rewritten as:

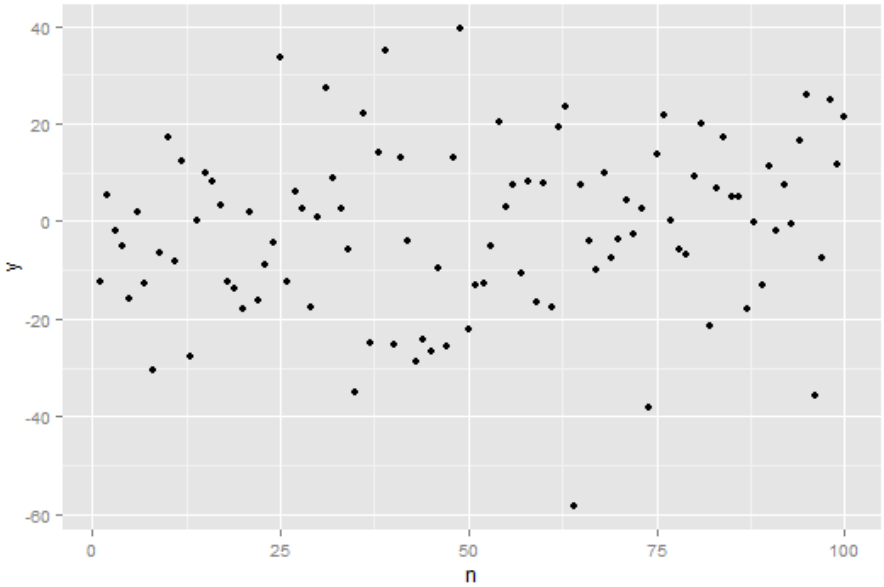
$$y_{i,t} = \mu_i + \beta' x_{i,t}(\gamma) + \varepsilon_{i,t} \tag{2}$$

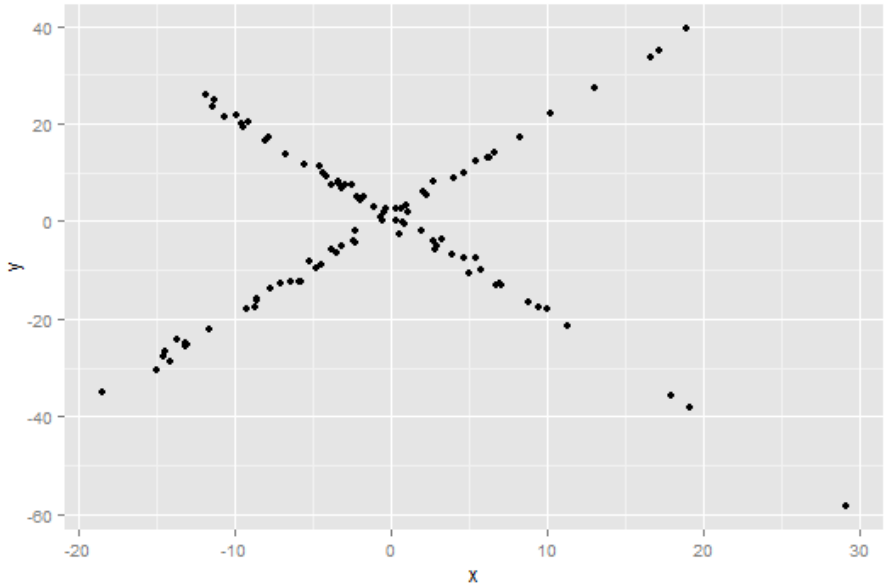
Simulation

- First of all, we need to generate a set of original data for the following simulation.

$$y_n = \begin{cases} 1 + 2x_n + \varepsilon_n & 0 < n \leq 50 \\ 1 - 2x_n + \varepsilon_n & 50 < n \leq 100 \end{cases}$$

And the relationship between y and n and y and x has been plotted as follows:





Simulation

- If we do know where the threshold is and we choose to ignore it, what may happen?
- Regression without considering the threshold:

```
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-61.257 -10.452   1.005  10.659  38.196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4012     1.7574  -0.797    0.427
x             0.1474     0.2017   0.731    0.466

Residual standard error: 17.47 on 98 degrees of freedom
Multiple R-squared:  0.005424,  Adjusted R-squared:  -0.004725
F-statistic: 0.5344 on 1 and 98 DF,  p-value: 0.4665
```

Simulation

- Regressions with considering the threshold:

```
lm(formula = y ~ x, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9905 -0.6041 -0.1065  0.6652  2.1689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0413     0.1366   7.623 8.23e-10 ***
x             2.0030     0.0150 133.514  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9351 on 48 degrees of freedom
Multiple R-squared:  0.9973,    Adjusted R-squared:  0.9973
F-statistic: 1.783e+04 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
lm(formula = y ~ x, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.45155 -0.52374  0.05102  0.41688  2.12632

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  1.15231    0.12790    9.009 6.87e-12 ***
x           -2.01689    0.01541 -130.914  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.9036 on 48 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9971
F-statistic: 1.714e+04 on 1 and 48 DF,  p-value: < 2.2e-16
```

- What if we do not know where the threshold point is in advance?

Model Estimation

- The method used to eliminate the individual effect μ_i is to remove individual-specific means.

$$\bar{y_i} = \mu_i + \beta' \bar{x_i}(\gamma) + \bar{\varepsilon_i} \tag{3}$$

- Taking the difference between (2) and (3) yields

$$y^*_{i,t} = \beta' x^*_{i,t}(\gamma) + \varepsilon^*_{i,t} \tag{4}$$

- Letting $Y^*, X^*(\gamma)$ and ε^* denote the data stacked over all individuals, then equation (4) equals to

$$Y^* = X^*(\gamma)\beta + \varepsilon^* \tag{5}$$

Model Estimation

- For any given γ, the slope coefficient β can be estimated by OLS, which is

$$\hat{\beta}(\gamma) = (X^*(\gamma)'X^*(\gamma))^{-1}X^*(\gamma)'Y^* \tag{6}$$

- The vector of regression residuals is

$$\hat{\varepsilon}^*(\gamma) = Y^* - X^*(\gamma)\hat{\beta}(\gamma) \tag{7}$$

- The sum of squared errors is

$$\begin{aligned} S_1(\gamma) &= \hat{\varepsilon}^*(\gamma)'\hat{\varepsilon}^*(\gamma) \\ &= Y^{*\prime}(I - X^*(\gamma)'(X^*(\gamma)'X^*(\gamma))^{-1})Y^* \end{aligned} \tag{8}$$

Model Estimation

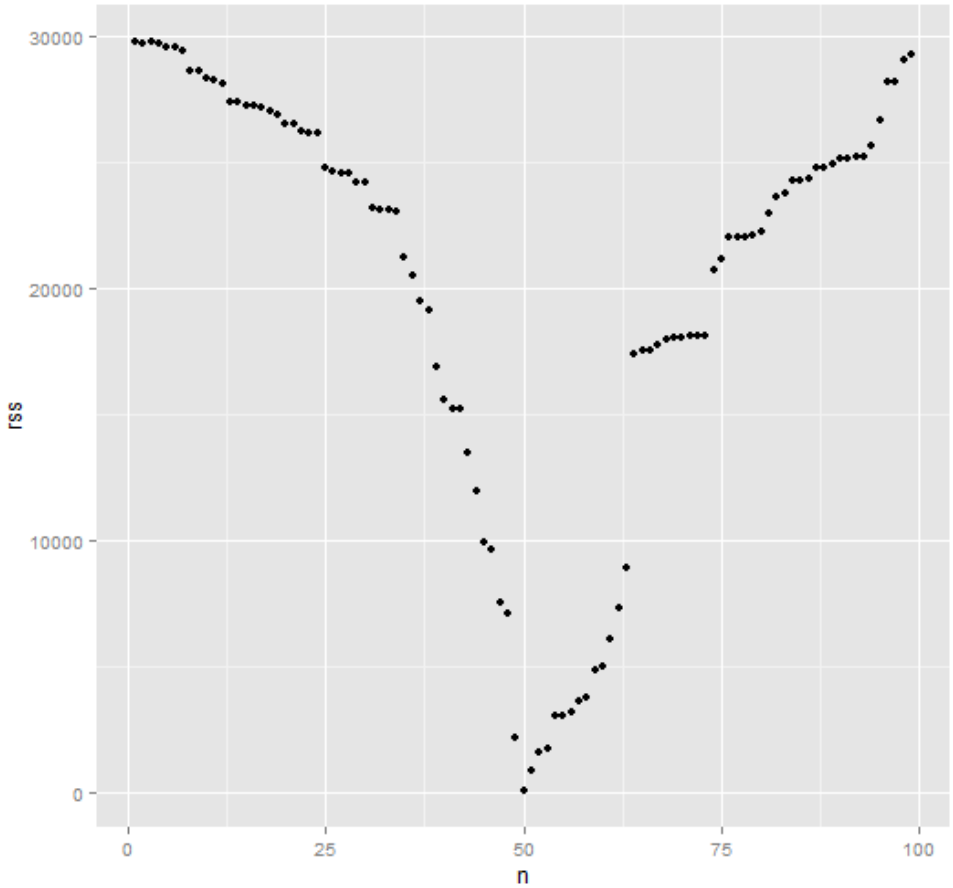
- Chan(1993) and Hansen(1999) recommend estimation of γ by least squares. This is easiest to achieve by minimization of the concentrated sum of squared errors (7). Hence the least squares estimators of γ is

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}}\, S_1(\gamma) \tag{9}$$

  In this situation, the γ we get is the threshold value we are seeking.
- Nowadays, with the dramatical increase of compute ability, we can easily discern the threshold point by using powerful computer. In the following part, I will try to illustrate how to use R to conduct the above process.

Simulation

- The result of loop can be plotted as follows. We can easily find out when n=50 ,the rss reachs its minimum value, which is the same point we set in advance. Hence, we successfully discern the threshold value.



Model With More Than One Threshold

- The above simple model has a single threshold. In some applications there may be multiple thresholds. For example, the double threshold model takes the following form:

$$y_{it} = \mu_i + \beta_1' x_{it} I(q_{it} \leq \gamma_1) + \beta_2' x_{it} I(\gamma_1 < q_{it} \leq \gamma_2) + \beta_3' x_{it} I(\gamma_2 < q_{it}) + e_{it} \tag{10}$$

- Two methods of discerning these thresholds
  - Similar with the method used in simple model, we can conduct n^2 regressions and find out the two thresholds.
  - Use the method proposed by Hansen(1999).

Multiple Thresholds Estimation

- Let $S_1(\gamma)$ be the single threshold sum of squared errors as defined in equation(8) and let $\hat{\gamma}_1$ be the threshold estimate which minimizes $S_1(\gamma)$.
- Fixing the first-stage estimate $\hat{\gamma}_1$, the second-stage criterion is

$$S_2^r(\gamma_2) = \begin{cases} S(\hat{\gamma}_1, \gamma_2) & if \hat{\gamma}_1 < \gamma_2 \\ S(\gamma_2, \hat{\gamma}_1) & if \gamma_2 < \hat{\gamma}_1 \end{cases} \tag{11}$$

and the second-stage threshold estimate is

$$\hat{\gamma}_2^r = \underset{\gamma_2}{\operatorname{argmin}}\, S_2^r(\gamma_2) \tag{12}$$

Multiple Thresholds Estimation

- Bai(1997) has shown that $\hat{\gamma}_2^r$ is asymptotically efficient, but $\hat{\gamma}_1$ is not. The asymptotic efficiency of $\hat{\gamma}_2^r$ suggests that $\hat{\gamma}_1$ can be improved by a third-stage estimation. He suggested the following refinement estimator.
- Fixing the second-stage estimate $\hat{\gamma}_2^r$, define the refinement criterion as follow:

$$S_1^r(\gamma_1) = \begin{cases} S(\gamma_1, \hat{\gamma}_2^r) & if \gamma_1 < \hat{\gamma}_2^r \\ S(\hat{\gamma}_2^r, \gamma_1) & if \hat{\gamma}_2^r < \gamma_1 \end{cases} \tag{13}$$

and the refinement estimate is

$$\hat{\gamma}_1^r = \underset{\gamma_1}{\operatorname{argmin}}\, S_1^r(\gamma_1) \tag{14}$$

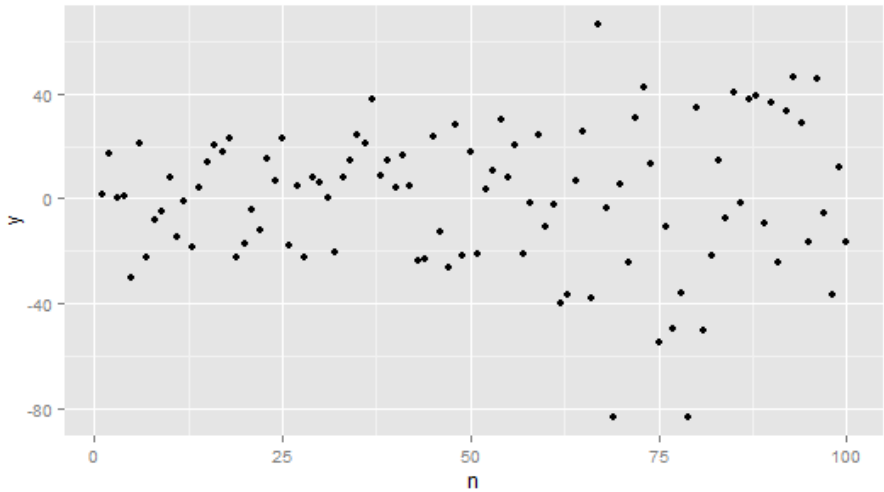
- In the following part, I will try to illustrate how to use R to conduct the above process.

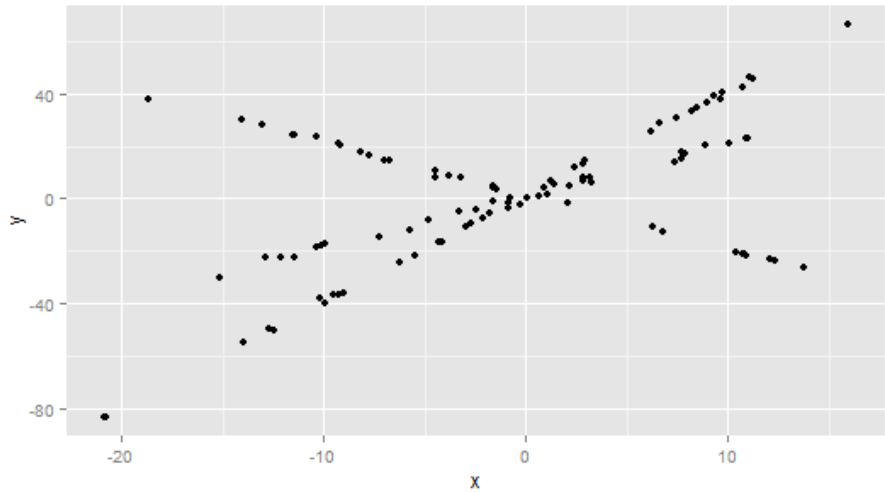
simulation

- First of all, we need to generate a set of original data for the following simulation.

$$y_n = \begin{cases} 1 + 2x_n + \varepsilon_n & 0 < n \leq 30 \\ 1 - 2x_n + \varepsilon_n & 30 < n \leq 60 \\ 1 + 4x_n + \varepsilon_n & 60 < n \leq 100 \end{cases}$$

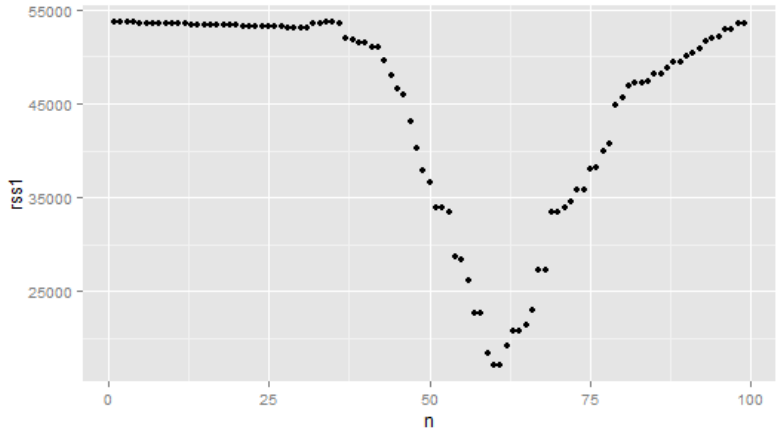
And the relationship between y and n and y and x has been plotted as follows:



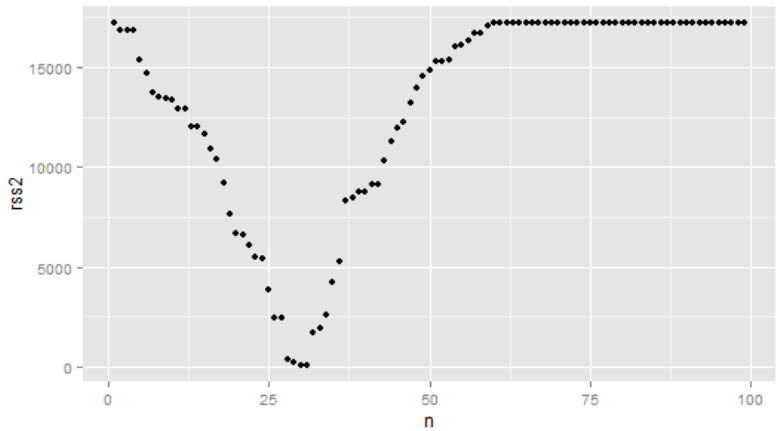


simulation

- We conduct the first loop to find out the first threshold value.

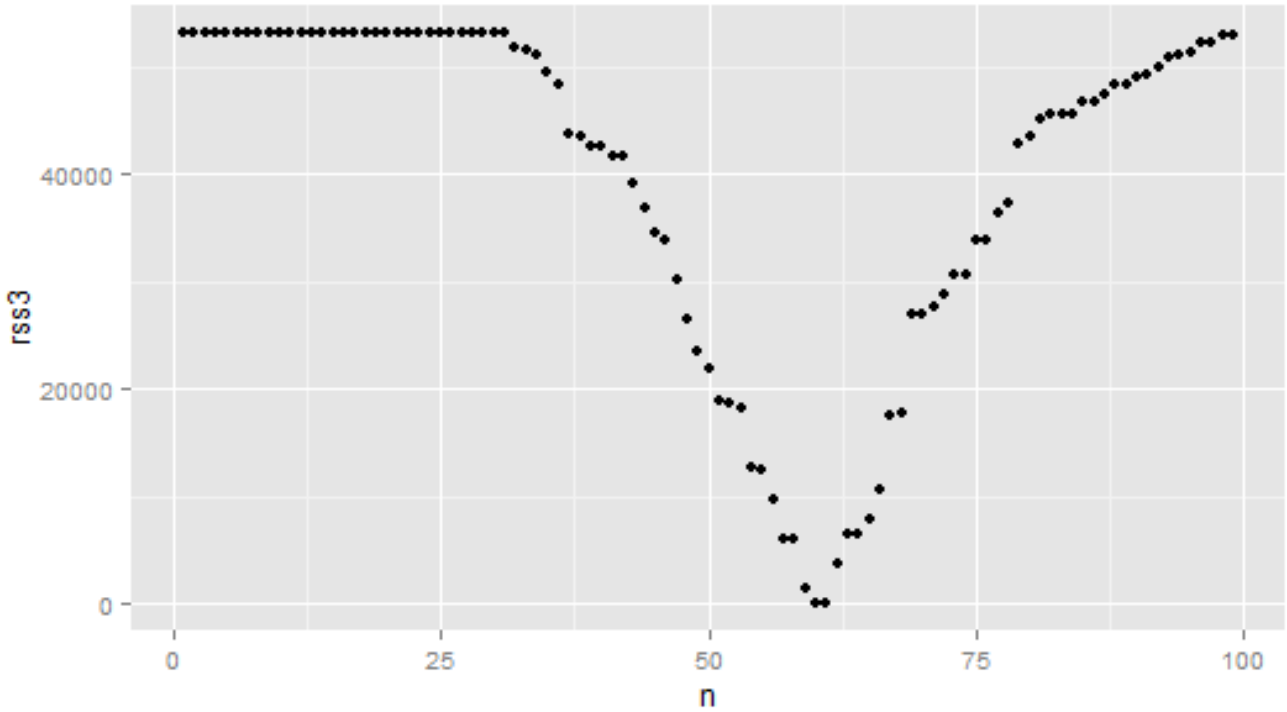


- Then we conduct the second loop to find out the second first threshold value.



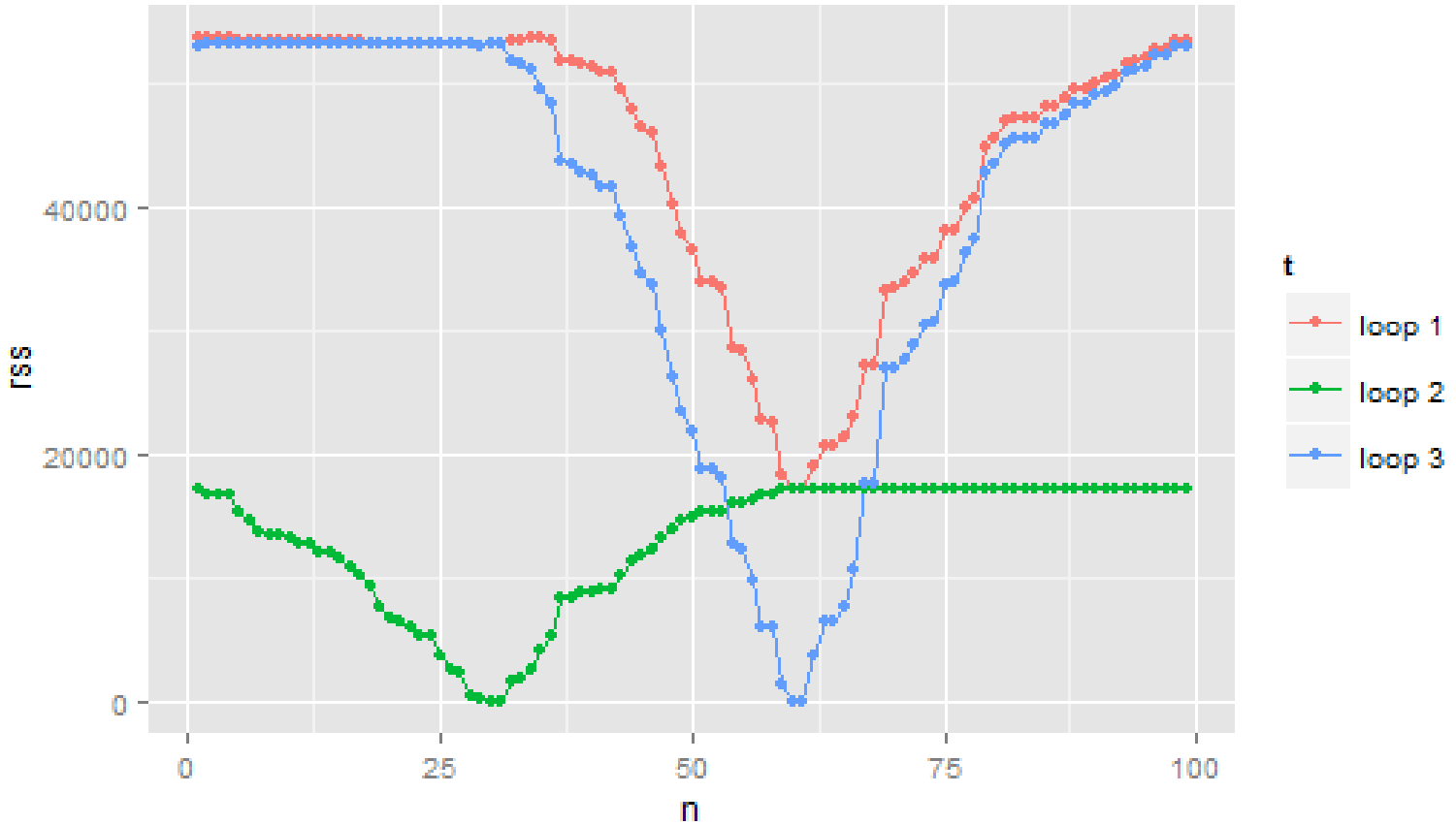
simulation

- Fix the second threshold value and search the first threshold value again.



simulation

- I put the three plot figures into one picture to help you scrutinize those subtle relationships among them.



Does Threshold Effect Really Exist?

- It is important to determine whether the threshold effect is statistically significant.

$$\text{H0: } \beta_1 = \beta_2 \qquad \text{H1:} \beta_1 \neq \beta_2$$

- The likelihood ratio test of H0 is based on:

$$F_1 = \frac{(S_0 - S_1(\hat{\gamma}))}{\hat{\sigma}^2} \tag{15}$$

- Under H0 the threshold γ is not identified, so the asymptotic distribution of F_1 is non-standard. This is typically called the "Davies' Problem". Hansen(1996) suggested a bootstrap to simulate the asymptotic distribution of the LR test.

Does Threshold Effect Really Exist?

- Procedure of Bootstrap:
  - Treat the regressors x_{it} and threshold variable q_{it} as given, holding their values fixed in repeated bootstrap sample. Take the regression residuals $\hat{e}^*_{it}$, and group them by individual: $\hat{e}^*_i = (\hat{e}^*_{i1}, \hat{e}^*_{i2}, ..., \hat{e}^*_{iT})$. Treat the sample $\{\hat{e}^*_1, \hat{e}^*_2 ..., \hat{e}^*_n\}$ as the empirical distribution to be used for bootstrapping.
  - Draw (with replacement) a sample of size n from the empirical distribution and use these errors to create a bootstrap sample under H0, i.e. $Y_{bs} = X^*(\gamma)\beta + e_{bs}$.
  - Using the bootstrap sample, estimate the model of null hypothesis: $y^*_{it} = \beta'_1 x^*_{it} + e^*_{it}$ and the model of alternative hypothesis: $y^*_{it} = \beta'_1 x^*_{it}(\gamma) + e^*_{it}$ and calculate the bootstrap value of the likelihood ratio statistic F_1.
  - Repeat this procedure a large number of times (99,199,1999...) and calculate the percentage of draws for which the simulated statistic exceeds the actual. This is the bootstrap estimate of the asymptotic p-value for F_1 under H0.

Does Threshold Effect Really Exist?

- In the following part, I will try to illustrate how to use R to conduct the above process.
- Itroduction of the R package, "boot".

```
#bootstrap
breg1=lm(y~x,data)
s0=sum(residuals(breg1)^2)
library(boot)
fvalue=function(formula1,data,indices){
  d=data[indices,]
  bregloop=list()
  brss=array()
  for (i in 1:99)
  {
    dum=d$x
    dum[(i+1):100]=0
    bregloop[[i]]=lm(formula1,data=d)
    brss[i]=sum(residuals(bregloop[[i]])^2)
  }
  a=which.min(brss)
  dum=d$x
  dum[(a+1):100]=0
  breg2=lm(formula1,data=d)
  s1=sum(residuals(breg2)^2)
  f=(s0/s1-1)*(100-1)
  return(f)
}

result=boot(data=data,fvalue,R=99,formula1=y~x+dum)
f=result$t
```

How to construct the confidence intervals for γ?

- When there is a threshold effect $(\beta_1 \neq \beta_2)$ Chan(1993) and Hansen(1999) have shown that $\hat{\gamma}$ is consistent for γ_0, which is the true value of γ and that the asymptotic distribution is highly non-standard.
- Hansen(1999) argues that the best way to form confidence intervals for γ is to form the "no-rejection region" using the likelihood ratio statistic for tests on γ.
- The procedure of confidence intervals construction of γ will be illustrated as follow:

How to construct the confidence intervals for γ?

- To test the hypothesis H0:$\gamma = \gamma_0$, the likelihood ratio test is to reject for large values of $LR_1(\gamma_0)$.

$$LR_1(\gamma) = \frac{(S_1(\gamma) - S_1(\hat{\gamma}))}{\hat{\sigma}^2} \tag{16}$$

  Note that the statistic LR_1 is testing a different hypothesis from the statistic F_1 introduced in the previous section. $LR_1(\gamma_0)$ is testing H0:$\gamma = \gamma_0$, while F_1 is testing $\beta_1 = \beta_2$.

- Hansen (1999) proposes to use the following equation to calculate critical values:

$$c(\alpha) = -2log(1 - \sqrt{1 - \alpha}) \tag{17}$$

  A test of H0: $\gamma = \gamma_0$ rejects at the asymptotic level α if $LR_1(\gamma_0)$ exceeds $c(\alpha)$.

An Example: Government Size and Economic Growth

- We use the paper, *Government size and economic growth in Taiwan: A threshold regression approach (Sheng-Tung Chen, Chien-Chiang Lee (2005))* to illustrate how to conduct your own Economic research using threshold regression method.

ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Policy Modeling 27 (2005) 1051–1066

Journal of Policy Modeling

Government size and economic growth in Taiwan: A threshold regression approach

Sheng-Tung Chen, Chien-Chiang Lee*

Department of Applied Economics, National Chung Hsing University, 250 Kuo-Kuang Rd. Taichung 402, Taiwan, ROC

Received 1 October 2004; received in revised form 1 March 2005; accepted 1 June 2005
Available online 8 August 2005

Introduction

- Does expanding government size promote economic growth? What government size is optimum?

Table 1
Literature discussing the relationship of government size and economic growth

Authors	Relationship of government size and economic growth	Empirical method	Subject	Explanation
Landau (1983)	Negative	OLS	96 developed countries	Classify government expenditure
Engen and Skinner (1991)	Negative	2SLS	107 countries	
Fölster and Henrekson (2001)	Negative	OLS	23 OECD countries and 7 developing countries	
Dar and AmirKhalkhali (2002)	Negative	Random coefficient model	19 OECD countries	
Ram (1986)	Positive	OLS	115 countries	Discuss the difference while the time is divided
Kormendi and Meguire (1986)	Positive	OLS	47 countries	The government size indicator is the average growth rate of total government expenditure/total private consumption expenditure

- For the inconsistency of the above result, Vedder and Gallaway (1998) and Sheehey (1993) point out that the reason is that government size and economic growth exist under a non-linear relationship.

Introduction

- Armey (1995) implements the Laffer curve to present the relationship between government size and economic growth.

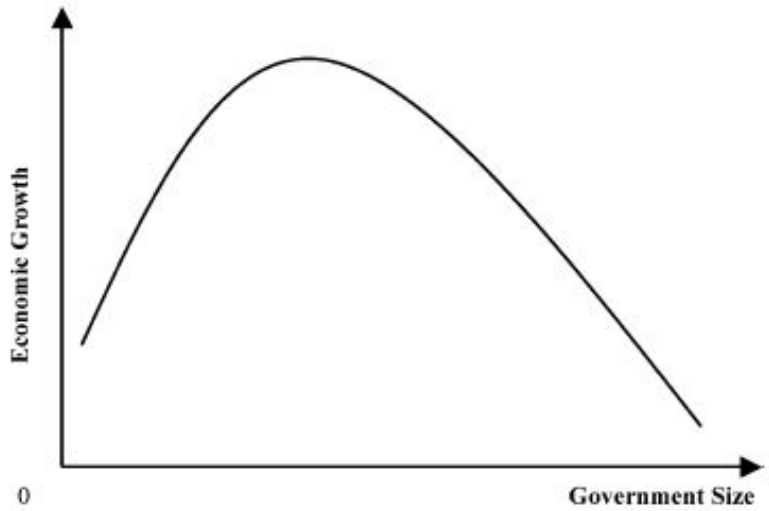


Fig. 1. Armey curve.

- This paper uses the threshold regression model to test whether the Armey curve exists in Taiwan or not.

Model

- The empirical regression equation is written in this paper as follows:

$$\dot{Y}_t = \alpha_0 + \alpha \frac{I_t}{Y_t} + \alpha_2 \dot{L}_t + \alpha_3 \dot{G}_t \frac{G_t}{Y_t} + u_t^* \tag{18}$$

  This equation shows that the variables which affect economic growth ($\dot{Y}$) include the investment rate ($\frac{I}{Y}$), labor force growth ($\dot{L}$), and the multiple effects of the growth of government expenditure ($\dot{G}$) and government size ($\frac{G}{Y}$).
- This indicates that the government sector has a reciprocal effect on economic growth through two ways: one is the direct contribution of the government sector and the other is the indirect effect of government sector through the non-government sector.

Model

- Equation (18) is the traditional linear economic growth model, now we alter the linear model into the two-regime threshold regression model. The model can be shown as follows:

$$\dot{Y}_t = (\delta_{10} + \delta_{11}(\frac{I_t}{Y_t}) + \delta_{12}\dot{L}_t + \delta_{13}\dot{G}_t(\frac{G_t}{Y_t}))I[q_t \leq \gamma]$$
$$+(\delta_{20} + \delta_{21}(\frac{I_t}{Y_t}) + \delta_{22}\dot{L}_t + \delta_{23}\dot{G}_t(\frac{G_t}{Y_t}))I[q_t > \gamma] + u_t^* \qquad (19)$$

- The threshold value $gamma$ can be found by estimating equation (19) through finding the minimum one of the SSEs in a re-order threshold variable.
- Before finding out the threshold value, we need to test if there real exists thershold(s). $H_0 : \delta_{1i} = \delta_{2i} \qquad i = 0, 1, 2, 3$

Results

Threshold test

Threshold variable	Total government expenditure/GDP	Government investment expenditure/GDP	Government consumption expenditure/GDP
F_1 value of one threshold test	15.230** (0.009)	12.752** (0.036)	19.079** (0.000)
F_2 value of two threshold tests	8.686 (0.263)	5.915 (0.642)	4.160 (0.936)
Threshold regime (%)	22.839	7.302	14.967
95% confidence interval	[22.655%, 25.452%]	[7.000%, 10.503%]	[12.626%, 17.108%]

Values given in parenthesis denotes bootstrap p-value. Values given in square parenthesis means confidence interval.

** Indicates significance at a 5% level.

- Using the bootstrap model proposed by Hansen, the authors find out that there is only one threshold existing, and the threshold value for different indexes of government size are separately 22.839, 7.302 and 14.967. Now we can classify the data into two regimes to analyze the effect of government size on economic growth.

Results

Table 7
The regression results of government size (total government expenditure/GDP) and economic growth

Dependent variables	Linear model	Government size is small	Government size is large
Threshold value (%)		≤22.839	>22.839
Interception	2.721 (6.209)**	1.295 (1.832)*	3.188 (6.124)**
$\left(\frac{I}{Y}\right)$	0.286 (2.224)**	0.279 (1.086)	0.337 (2.478)**
$\dot{L}$	1.889 (8.481)**	2.351 (10.392)**	1.381 (3.773)**
$\dot{G}\left(\frac{G}{Y}\right)$	−0.047 (−3.031)**	0.675 (4.447)**	−0.036 (−2.158)**
$\overline{R}^2$	0.565	0.902	0.386
Number of Samples	99	20	79

Threshold variable is total government expenditure/GDP. Values given in parenthesis denote *t*-value.
* Indicate significance at 10% level.
** Indicate significance at 5% level.

Table 8
The regression results of government size (government investment expenditure/GDP) and economic growth

Dependent variables	Linear model	Government size is small	Government size is large
Threshold value (%)		≤7.302	>7.302
Interception	2.652 (6.071)**	1.614 (1.446)	3.101 (5.925)**
$\left(\frac{I}{Y}\right)$	0.271 (2.117)**	0.210 (0.593)	0.378 (2.436)**
$\dot{L}$	1.892 (8.436)**	2.332 (6.829)**	1.297 (3.182)**
$\dot{G}\left(\frac{G}{Y}\right)$	−0.026 (−4.859)**	0.815 (2.944)**	−0.016 (−1.920)*
$\overline{R}^2$	0.564	0.710	0.418
Number of samples	99	23	76

Table 9
The regression results of government size (government consumption expenditure/GDP) and economic growth

Dependent variables	Linear model	Government size is small	Government size is large
Threshold value (%)		≤14.967	>14.967
Interception	2.494 (5.271)**	1.702 (3.553)**	4.724 (5.441)**
$\left(\frac{I}{Y}\right)$	0.245 (1.828)*	0.384 (1.984)*	0.036 (0.206)**
$\dot{L}$	1.874 (8.330)**	1.911 (6.186)**	1.487 (4.113)**
$\dot{G}\left(\frac{G}{Y}\right)$	0.102 (0.428)	1.039 (3.254)**	−0.419 (−1.458)
$\overline{R}^2$	0.558	0.783	0.361
Number of samples	99	43	56

Threshold variable is government consumption expenditure/GDP. Values given in parenthesis denotes *t*-value.
* Indicate significance at 10% level
** Indicate significance at 5% level.

Thank You

All the code can be downloaded from my homepage: www.yzc.me