SO SÁNH TÍNH HIỆU QUẢ MỘT SỐ PHƯƠNG PHÁP ƯỚC LƯỢNG ĐIỂM CHUYỂN TRONG MÔ HÌNH HỒI QUY TUYẾN TÍNH CÓ GÃY

Nguyễn Thị Quyên
Khoa Toán và Khoa học tự nhiên
Email: quyennt@dhhp.edu.vn

Ngày nhận bài: 04/10/2019
Ngày PB đánh giá: 25/10/2019
Ngày duyệt đăng: 25/11/2019

TÓM TẮT: Trong bài báo này, thông qua nghiên cứu mô phỏng bằng phần mềm thống kê R, tác giả so sánh hiệu quả ước lượng thời điểm chuyển, điểm chuyển của mô hình hồi quy gãy khúc liên tục của ba phương pháp: dựa vào ước lượng bình phương cực tiểu SEG, dựa vào tỉ số hợp lý thực nghiệm ELR và dựa vào phần dư dịch chuyển tham số CĐBP. Các mô phỏng được thực hiện với một số phân bố của biến độc lập và sai số. Kết quả chỉ ra rằng khi ước lượng thời điểm chuyển thì phương pháp CĐBP tỏ ra hiệu quả hơn phương pháp ELR, nhưng khi ước lượng điểm chuyển, dao độ hệ số góc khá lớn thì phương pháp SEG hiệu quả hơn phương pháp CĐBP.

Từ khóa**:** *điểm chuyển, hiệu quả, mô hình hồi quy gãy, thời điểm chuyển, ước lượng.*

COMPARING THE EFFICIENCY OF SOME METHODS ESTIMATING THE CHANGE-POINT OF THE SEGMENTED REGRESSION

ABSTRACT: In this paper, by simulating through statistical software R, we compare the efficiency of estimating the time of the change and the change-point of a continuous segmented regression model of three methods: the Ordinary least squared (SEG); empirical likelihood ratios (ELR) and the shift of parameter estimated residuals (CDBP). Simulation studies are performed with some distributions of independent var and error. The result shows that when estimating the time of the change, the CDBP method is more effective than the ELR one, but when estimating the change- point, with the change of slope is quite large, the SEG method is more effective than the CDBP one.

Keywords**:** *change-point, efficiency, segmented regression model, the time of the change, estimated.*

1. GIỚI THIỆU

Nhiều quá trình trong thực tế tuân theo mô hình hồi quy tuyến tính hai pha, ở đó các tham số điều khiển mô hình giữ nguyên giá trị trong pha đầu, tại một thời điểm nào đó nó chuyển sang giá trị khác và giữ nguyên trong pha còn lại. Việc nghiên cứu mô hình có thay

đổi trạng thái như vậy - còn gọi là mô hình điểm chuyển - đã được phát triển hơn nửa thế kỷ qua và đạt được những thành tựu rực rỡ, được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau. Trong kinh tế, người ta thấy mô hình điểm chuyển bội là phù hợp khi nghiên cứu mối quan hệ giữa lãi suất (interets rate) đối với thay đổi lãi suất chiết khấu (discount rate) quy định bởi FED. Sử dụng mô hình ARCH để nghiên cứu chuỗi thời gian trong miền tần số, người ta đã phát hiện ra sự chuyển đổi của chuỗi thời gian chỉ số chứng khoán, cũng như thị trường ngoại hối liên hệ mật thiết với khủng hoảng tài chính ở châu Á và Liên Xô. Việc nghiên cứu mô hình điểm chuyển là cần thiết và liên tục được phát triển trong những năm gần đây. Cụ thể, xét mô hình

$$y_i = \begin{cases} \alpha_0 + \alpha_1 x_i + \varepsilon_i & khi \;\; 1 \le i \le k^*, \\ \beta_0 + \beta_1 x_i + \varepsilon_i & khi \;\; k^* \le i \le n \end{cases} \qquad (1)$$

trong đó $a \le x_1 < ... < x_n \le b$, các sai số $\{\varepsilon_i\}$ là ngẫu nhiên, $\alpha_0, \alpha_1, \beta_0, \beta_1, k^*$ chưa biết.

Nếu $\alpha_0 = \beta_0$ và $\alpha_1 = \beta_1$ thì mô hình (1) được gọi là không có chuyển. Ngược lại, nếu ít nhất một trong hai đẳng thức này không xảy ra, mô hình được gọi là có chuyển và k^* được gọi là thời điểm chuyển. Đối với mô hình có chuyển, $\alpha_1 \ne \beta_1$, và hai đường thẳng $y = \alpha_0 + \alpha_1 x$ và $y = \beta_0 + \beta_1 x$ cắt nhau tại điểm có hoành độ $x = \tau$ thì τ được gọi là điểm chuyển. Trong trường hợp $\tau \in [x_{k^*}, x_{k^*+1})$ thì hàm mô hình được gọi là gãy khúc liên tục, mô hình được gọi là liên tục. Khi đó, (1) được viết lại dưới dạng $y_i = f(x_i) + \varepsilon_i$, trong đó

$$f(x) = \alpha_0 + \alpha_1 x + h(x-\tau) I(x > \tau) = \alpha_0 + \alpha_1 x + h(x-\tau)^+ . \qquad (2)$$

Để ước lượng thời điểm chuyển k^* và điểm chuyển τ của mô hình (1), (2), nhiều tác giả đã đưa ra các phương pháp khác nhau; tuy nhiên hiệu quả của các phương pháp chưa được đánh giá, xem xét, bởi khi xem xét một hàm mô hình, nếu nó thỏa mãn các giả thiết của các phương pháp thì ta cần cân nhắc lựa chọn phương pháp nào để có được kết quả tốt nhất – ước lượng được thời điểm chuyển, và điểm chuyển tốt nhất. Trong bài báo này, tác giả sử dụng phần mềm thống kê R để so sánh hiệu quả ước lượng thời điểm chuyển, điểm chuyển của ba phương pháp: phương pháp SEG, phương pháp tỉ số hợp lý (ELR) và phương pháp cực đại tổng bình phương phần dư dịch chuyển (CĐBP) đối với hàm mô hình tuyến tính đơn, gãy khúc liên tục.

2. MỘT SỐ PHƯƠNG PHÁP ƯỚC LƯỢNG ĐIỂM CHUYỂN

2.1 Phương pháp bình phương cực tiểu (SEG)

Muggeo, V.M.R. [3] đã đưa ra phương pháp để ước lượng các điểm chuyển liên tiếp của mô hình hồi quy gãy khúc liên tiếp. Trong bài này, chúng ta quan tâm tới việc nghiên cứu điểm chuyển của mô hình với hàm gãy liên tục tại một điểm τ nào đó:

$$f(x_i) = \alpha_0 + \alpha_1 x_i + h(x_i - \tau)^+$$

Sử dụng khai triển Taylor bậc 1 tại lân cận điểm τ_0:

$$(x_i - \tau)^+ \approx (x_i - \tau_0)^+ + (\tau - \tau_0).(-1).\mathrm{I}_{(x_i > \tau_0)} .$$

Khi đó, hàm mô hình (1) được xấp xỉ

$$f(\mathrm{x}_i) \approx \alpha_0 + \alpha_1 x_i + h\left((x_i - \tau_0)^+ + (\tau - \tau_0).(-1).\mathrm{I}_{(x_i > \tau_0)}\right)$$

$$= \alpha_0 + \alpha_1 x_i + h(x_i - \tau_0)^+ - \gamma\, \mathrm{I}_{(x_i > \tau_0)}, \qquad (3)$$

với $\gamma = h(\tau - \tau_0)$.

Sử dụng phương pháp bình phương nhỏ nhất (hoặc hàm hợp lý), ước lượng được các tham số $\hat{\alpha}_0, \hat{\alpha}_1, \hat{h}, \hat{\gamma}$ của hàm mô hình (3). Khi đó, điểm chuyển τ được ước lượng bởi $\hat{\tau} = \tau_0 + \dfrac{\hat{\gamma}}{\hat{h}}$.

Cách ước lượng điểm chuyển được Muggeo, V.M.R. xây dựng thông qua các bước sau:

(i) Chọn một điểm chuyển ban đầu là τ_0.

(ii) Với điểm chuyển τ_0 đã chọn, bằng cách sử dụng phương pháp bình phương nhỏ nhất, ước lượng được các tham số, từ đó cập nhật điểm chuyển ước lượng mới bởi: $\hat{\tau}_1 = \tau_0 + \hat{\gamma} / \hat{h}.$

(iii) Tiếp tục lặp lại bước (i) và (ii), thu được dãy các điểm chuyển ước lượng $\{\hat{\tau}_i\}$. Khi dãy điểm chuyển có xu thế hội tụ thì dừng lại và giới hạn của dãy được dùng để ước lượng điểm chuyển thực của hàm mô hình.

2.2 Phương pháp tỉ số hợp lý thực nghiệm (ELR)

Xét mô hình gãy liên tục với hàm mô hình (2), ở đó $\{\varepsilon_i\}$ là dãy các sai số ngẫu nhiên độc lập với kì vọng không. Liu Z. và Qian L. [4] đã đưa ra phương pháp ước lượng thời điểm chuyển bằng việc sử dụng phần dư dịch chuyển. Cụ thể, để ước lượng sai số đoạn đầu, các tác giả đã sử dụng tham số ước lượng của đoạn sau và ngược lại như sau:

$$\tilde{e}_i(k) = \begin{cases} y_i - [\hat{\beta}_0(k) + \hat{\beta}_1(k)\, x_i], \; i = 1, \ldots, k; \\ y_i - [\hat{\alpha}_0(k) + \hat{\alpha}_1(k)\, x_i], \; i = k+1, \ldots, n. \end{cases} \qquad (4)$$

Nhận xét rằng, nếu mô hình không có chuyển thì $E[\tilde{e}_i(k)] = 0, \forall k$. Với mỗi k, đặt $\omega_i(k)$ là khối lượng xác suất tại giá trị $\tilde{e}_i(k)$ với điều kiện $\sum_{i=1}^{n} \omega_i(k) = 1$. Khi đó, theo Owen A. [2], mô hình được khẳng định có chuyển nếu tỉ số thực nghiệm hợp lí (ELR) sau đây đủ nhỏ.

$$\Re(k) = \left\{ \sup \prod_{i=1}^{n} n\omega_i(k) \,\middle|\, \sum_{i=1}^{n} \omega_1(k)\tilde{e}_i(k) = 0, \omega_i(k) \geq 0, \sum_{i=1}^{n} \omega_i(k) = 1 \right\}$$

Liu Z. và Qian L. đã đề xuất thống kê $Z_n = \sqrt{M_n} = \sqrt{\max_{3 \leq k \leq n-3} \{-2 \ln \Re(k)\}}$ và thống kê chặt cụt $Z_n' = \sqrt{M_n'} = \sqrt{\max_{L \leq k \leq U} \{-2 \ln \Re(k)\}}$, với $L = \left[\ln n\right]^2$, $U = n - L$, trong đó $[x]$ là số nguyên nhỏ nhất lớn hơn x, để kiểm định giả thuyết về sự tồn tại điểm chuyển. Như trong [3],

mô hình sẽ được coi là có chuyển khi Z_n hoặc $Z_n^{'}$ đủ lớn. Khi đó, thời điểm chuyển được ước lượng bởi:

$$\hat{k} = \min\{k : M_n = -2\ln \Re(k)\}$$

Khi đó Liu Z. đề xuất một thuật toán bao gồm sáu bước sau:

(i) Với mỗi k cố định, $k = 3,4,..., n-3$, chia dữ liệu ra thành 2 nhóm gồm hai nhóm: nhóm pha trái $\{(x_i, y_i)\}_{i=1}^{k}$ và nhóm pha phải $\{(x_i, y_i)\}_{i=k+1}^{n}$.

(ii) Ước lượng hệ số $\hat{\alpha}_0(k), \hat{\alpha}_1(k)$ từ nhóm pha trái và $\hat{\beta}_0(k), \hat{\beta}_1(k)$ từ nhóm pha phải. Nếu $\hat{\tau}(k) = \dfrac{\hat{\beta}_0(k) - \hat{\alpha}_0(k)}{\hat{\alpha}_1(k) - \hat{\beta}_1(k)} \notin [x_k, x_{k+1}]$, thực hiện bước (iii); nếu không, thực hiện bước (iv).

(iii) Đặt $\hat{\tau}(k) = x_k$, sau đó ước lượng OLS cho các tham số của hàm mô hình (2) với $\tau = \hat{\tau}(k)$ được các ước lượng $\hat{\alpha}_0(k)$, $\hat{\alpha}_1(k)$ và $\hat{h}(k)$. Từ đó, ước lượng được các tham số

$\hat{\beta}_o(k) = \hat{\alpha}_0(k) - \hat{h}(k)x_k$ và $\hat{\beta}_1(k) = \hat{\alpha}_1(k) + \hat{h}(k)$.

(iv) Tính các phần dư dịch chuyển $\hat{e}_i(k) = y_i - \left[\hat{\beta}_0(k) + \hat{\beta}_1(k)x_i\right]$ với $i = 1,...,k$ và

$\hat{e}_i(k) = y_i - \left[\hat{\alpha}_0(k) + \hat{\alpha}_1(k)x_i\right]$ với $i = k+1,...,n$.

(v) Sử dụng $\{\hat{e}_i(k)\}_{i=1}^{n}$ là số liệu nhập vào hàm **el.test** trong phần mềm R để tính $-2\log\hat{\Re}(k)$.

(vi) Đối với mỗi k, lặp lại các bước từ (i) đến (v) để thu được dãy $\left\{-2\log\hat{\Re}(k)\right\}_{k=3}^{n-3}$. Xác định điểm cực đại $\hat{k}^* = \underset{3\le k\le n-3}{\operatorname{argmax}} \left\{-2\log\hat{\Re}(k)\right\}$ là ước lượng của thời điểm chuyển.

Điểm chuyển τ được ước lượng bởi:

$$\hat{\tau} = \begin{cases} \dfrac{\hat{\alpha}_1(\hat{k}^*) - \hat{\beta}_1(\hat{k}^*)}{\hat{\beta}_0(\hat{k}^*) - \hat{\alpha}_0(\hat{k}^*)}, \; khi \; \dfrac{\hat{\alpha}_1(\hat{k}^*) - \hat{\beta}_1(\hat{k}^*)}{\hat{\beta}_0(\hat{k}^*) - \hat{\alpha}_0(\hat{k}^*)} \in [x_{k^*}, x_{k^*+1}) \\ x_{\hat{k}^*}, \; khi \; \dfrac{\hat{\alpha}_1(\hat{k}^*) - \hat{\beta}_1(\hat{k}^*)}{\hat{\beta}_0(\hat{k}^*) - \hat{\alpha}_0(\hat{k}^*)} \notin [x_{k^*}, x_{k^*+1}) \end{cases}$$

2.3 Phương pháp cực đại tổng bình phương phần dư dịch chuyển (CĐBP)

Với cách xây dựng phần dư dịch chuyển như của Liu Z. và Qian L., trong [1], các tác giả đã đề xuất cách ước lượng thời điểm chuyển dựa vào phần dư dịch chuyển được xây dựng bởi:

$$\hat{k}_n = argmax \sum_{i=1}^{n} \hat{e}_i^2(k).$$

Dưới một số điều kiện, trong [1], các tác giả đã chứng minh được rằng $\frac{\hat{k}_n}{n} \xrightarrow{hcc} \tau$.

Việc xây dựng cách ước lượng điểm chuyển cho mô hình hồi quy liên tục (1), các hệ số thỏa mãn $\alpha_0 + \alpha_1\tau = \beta_0 + \beta_1\tau$ thông qua việc ước lượng thời điểm chuyển được thực hiện như sau.

Từ thời điểm chuyển ước lượng $\hat{k}_n$ - ta tính được các ước lượng cho các hệ số ở pha trái và pha phải $\hat{\alpha}_{0\hat{k}_n}$, $\hat{\alpha}_{1\hat{k}_n}$, $\hat{\beta}_{0\hat{k}_n}$, $\hat{\beta}_{0\hat{k}_n}$. Đường hồi quy mẫu giai đoạn đầu $y = \hat{\alpha}_{0\hat{k}_n} + \hat{\alpha}_{1\hat{k}_n} x$ và giai đoạn sau: $y = \hat{\beta}_{0\hat{k}_n} + \hat{\beta}_{1\hat{k}_n} x$ có hoành độ giao điểm $\tau^* = \frac{\hat{\alpha}_{0\hat{k}_n} - \hat{\beta}_{0\hat{k}_n}}{\hat{\beta}_{1\hat{k}_n} - \hat{\alpha}_{1\hat{k}_n}}$, không nhất thiết nằm trong $[\mathrm{x}_{\hat{k}_n}, x_{\hat{k}_n+1})$ nên điểm chuyển ước lượng $\hat{\tau}$ cho τ dựa vào τ^* như sau.

$$\hat{\tau} = x_{\hat{k}_n} I(\tau^* \le x_{\hat{k}_n}) + \tau^* I(x_{\hat{k}_n} < \tau^* < x_{\hat{k}_n+1}) + x_{\hat{k}_n+1} I(\tau^* \ge x_{\hat{k}_n+1}) \ .$$

3. SỬ DỤNG R ĐỂ SO SÁNH HIỆU QUẢ PHƯƠNG PHÁP ƯỚC LƯỢNG

Trong phần này, ta sẽ so sánh hiệu quả ước lượng thời điểm chuyển, điểm chuyển của mô hình hồi quy tuyến tính có gãy liên tục. Như trong [1], việc ước lượng sẽ không phụ thuộc vào hệ số của đoạn đầu mà chỉ phụ thuộc vào độ biến động của hệ số góc h trong mô hình hồi quy tuyến tính liên tục, tức là ta xét mô hình $y = h(x-\tau)^+ + \varepsilon$.

3.1 Hiệu quả ước lượng thời điểm chuyển

Để so sánh ba phương pháp trên, trước hết ta tiến hành mô phỏng để so sánh hiệu quả trong việc ước lượng thời điểm chuyển k^*. Trong phương pháp tỷ số thực nghiệm (ELR) của Liu Z., và phương pháp tổng bình phương cực đại phần dư dịch chuyển (CĐBP), việc ước lượng không cần giả thiết sai số có phân bố chuẩn. Vì vậy, ta sẽ xem xét hai phương pháp trong các trường hợp cụ thể sau: sai số được giả thiết phân bố chuẩn - cụ thể là $N(0, 0.5^2)$, và không có phân bố chuẩn - là $log\ N(0, 0.1)$. Ứng với mỗi sai số, ta sẽ xét mô hình có sự biến động của hệ số góc của đoạn sau so với đoạn trước là nhỏ, trung bình và lớn (tương ứng với $h = 0.5;\ 1;\ 2$). Theo Liu Z. [4], việc mô phỏng được thực hiện 1000 lần, và hiệu quả ước lượng được dựa vào tỷ số RF - là tỷ số giữa số lần thời điểm chuyển ước lượng rơi vào miền chấp nhận được trong tổng số 1000 lần chạy mô phỏng. Cũng theo Liu Z., ở đây ta sẽ xem xét thời điểm chuyển ước lượng từ $L = [lnn]^2$ đến $U = n - L$, với n là kích thước mẫu. Khi đó, dao độ $d = |\hat{k}^* - k^*|$ của thời điểm chuyển ước lượng và thời điểm chuyển thực là chấp nhận được nếu: $d \le D = \left[\frac{U-L}{A}\right]$.

Trước hết, ta xem xét mô hình với biến độc lập X được sinh ngẫu nhiên từ $N(0,1)$ thì miền giá trị của X gần như chắc chắn trong khoảng $(-3,\ 3)$ nên độ dài miền giá trị của X là $A=6$.

Với việc làm như của Liu Z. ta sẽ mô phỏng với kích thước mẫu $n = 50$ và thời điểm chuyển thực là $k^* = 25$ tức là số liệu được sinh từ mô hình $y_i = h(x_i - x_{25})^+ + e_i$, với x_i được sinh từ $N(0,1)$. Khi đó $L = 16,\ U = 34,\ D = 3$. Code R được dùng để so sánh như sau:

Code R để ước lượng thời điểm chuyển với phương pháp ELR.

```
Liu Z.=function (h,m); {p=numeric(m); for (t in 1:m); {n=50; x=(1:n)/n; c=rlnorm(n, meanlog=0, sdlog=0.1);
y=numeric(n);for (i in 1:25); {y[i]=c[i]}; for (i in 26:n);
 {y[i]=h*x[i]-h*x[25]+c[i]};
e=numeric(n);a=numeric(19);for(k in 16:34);
{a1= summary(lm(y[1:k]~x[1:k]))$coefficients[1];
a2= summary(lm(y[1:k]~x[1:k]))$coefficients[2];
b1= summary (lm(y[ (k+1):n]~x[(k+1):n]))$coefficients[1];
b2=summary(lm(y[(k+1):n]~x[(k+1):n]))$coefficients[2];
for (j in 1:k);{e[j]=y[j]-(b1+b2*x[j])};for (j in (k+1):n);
{e[j]=y[j]-(a1+a2*x[j])};
a[k-15]=el.test(e,mu=0)$`-2LLR`};p[t]=which.max(a)+15};
d=sqrt((p-25)*(p-25));g0= length(subset(d,d ==0));
g1=length( subset(d,d==1));g2= length( subset(d,d==2));
g3= length( subset(d,d==3));g4= length( subset(d,d==4));
g5= length( subset(d,d==5));g6= length( subset(d,d==6));
g7=length( subset(d,d>=7));print(g0+g1+g2+g3);
print(c(g0,g1,g2,g3,g4,g5,g6,g7))}
```

Code R để ước lượng thời điểm chuyển với phương pháp CĐBP

```
dx=function (h,m);{p=numeric(m);for (t in 1:m);{n=50; x=(1:n)/n;
c=rlnorm(n, meanlog=0, sdlog=0.1);y=numeric(n);for (i in 1:25); {y[i]=c[i]}; for (i in 26:n); {y[i]=h*x[i]-h*x[25]+c[i]};
e=numeric(n);a=numeric(19);for(k in 16:34);
{a1= summary (lm(y[1:k]~x[1:k]))$coefficients[1];
a2= summary (lm(y[1:k]~x[1:k]))$coefficients[2];
b1= summary (lm(y[ (k+1):n]~x[(k+1):n]))$coefficients[1];
b2= summary (lm(y[(k+1):n]~x[(k+1):n]))$coefficients[2];
for(j in 1:k);{e[j]=y[j]-(b1+b2*x[j])};for(j in (k+1):n);
{e[j]=y[j]-(a1+a2*x[j])};a[k-15]= sum(e^2)};p[t]=which.max(a)+15};
d=sqrt((p-25)*(p-25)) ;g0= length(subset(d,d ==0));
g1=length( subset(d,d==1));g2= length( subset(d,d==2));
g3= length( subset(d,d==3));g4= length( subset(d,d==4));
g5= length( subset(d,d==5));g6= length( subset(d,d==6));
```

```
g7= length( subset(d,d>=7));print(g0+g1+g2+g3);
print(c(g0,g1,g2,g3,g4,g5,g6,g7))}
```

Kết quả mô phỏng được thể hiện trong Bảng 1, 2 sau.

Bảng 1. Kết quả trong trường hợp sai số chuẩn $N(0, 0.5^2)$

d	$\varepsilon_i \sim N(0.5^2)$					
	h=0.5		h=2		h=4	
	CĐBP	ELR	CĐBP	ELR	CĐBP	ELR
0	*31*	19	*76*	73	*182*	99
1	*72*	45	*157*	118	*232*	163
2	*69*	44	*128*	102	*159*	119
3	*58*	55	*96*	100	*106*	98
4	*79*	44	*60*	94	*53*	101
5	*72*	56	*65*	80	*62*	78
6	*79*	70	*72*	95	*35*	73
>6	*540*	667	*346*	338	*171*	269
RF(‰)	***230***	163	***457***	393	***679***	479

Bảng 2. Kết quả trong trường hợp sai số $\varepsilon_i \sim \log N(0.1^2)$

d	$\varepsilon_i \sim \log N(0.1^2)$					
	h=0.5		h=2		h=4	
	CĐBP	ELR	CĐBP	ELR	CĐBP	ELR
0	*111*	76	*265*	181	*415*	234
1	*210*	155	*191*	245	*375*	340
2	*127*	101	*160*	154	*96*	151
3	*94*	103	*79*	104	*39*	60
4	*75*	100	*72*	74	*18*	59
5	*51*	72	*52*	61	*12*	37
6	*57*	67	*37*	38	*7*	34
>6	*275*	326	*144*	143	*38*	85
RF(‰)	*542*	435	*695*	684	*925*	785

Nhận xét: Kết quả mô phỏng với $N = 1000$ lần lặp lại bằng cách sư dụng phần mềm R được thể hiện ở Bảng 3 và 4. Chẳng hạn, đối với trường hợp $h = 2$, sai số $\varepsilon_i \sim N(0, 0.5^2)$ và thiết kế đều, theo phương pháp ELR trong 1000 lần mô phỏng có 76+157+128+96=457 lần giá trị d không vượt quá $D = 3$, như vậy $RF = 457$ ‰ cao hơn 393 ‰ (xem cột 4-5 Bảng 1).

So sánh kết qua ở các tình huống khác cặp cột khác (2-3, 6-7), hoặc trường hợp $\varepsilon_i \sim \log N(0.1^2)$, có thể kết luận rằng phương pháp tổng bình phương phần dư dịch chuyển tỏ ra hiệu quả hơn so với phương pháp ELR.

Bây giờ, ta sẽ mở rộng trong trường hợp biến giải thích được sinh từ phân bố đều trên $[-3, 3]$. Cụ thể với code ở trên, kết quả thu được được thể hiện trong bảng sau:

Bảng 3. Kết quả trong trường hợp thiết kế đều, sai số chuẩn $N(0,0.5^2)$

d	h=0.5		h=2		h=4	
	CĐBP	ELR	CĐBP	ELR	CĐBP	ELR
0	*50*	22	*215*	95	*415*	190
1	*56*	70	*191*	192	*375*	278
2	*56*	57	*110*	127	*96*	133
3	*60*	64	*79*	97	*39*	82
4	*59*	54	*72*	71	*18*	56
5	*74*	72	*52*	75	*12*	58
6	*88*	92	*37*	75	*7*	49
>6	*557*	569	*244*	268	*38*	154
RF(‰)	*222*	213	*595*	511	*925*	638

Bảng 4. Kết quả trong trường hợp thiết kế đều, sai số chuẩn $\log N(0.1^2)$

d	$\varepsilon_i \sim \log N(0.1^2)$					
	h=0.5		h=2		h=4	
	CĐBP	ELR	CĐBP	ELR	CĐBP	ELR
0	*267*	30	*528*	91	*516*	185
1	*251*	37	*468*	149	*484*	273
2	*138*	48	*4*	131	*0*	135
3	*81*	59	*0*	83	*0*	79
4	*53*	52	*0*	88	*0*	67
5	*32*	73	*0*	86	*0*	59
6	*22*	84	*0*	72	*0*	47
≥7	*156*	617	*0*	300	*0*	155
RF(‰)	*737*	174	*10000*	454	*1000*	672

Nhận xét: Dựa vào Bảng 3 và 4 ta cũng thấy ước lượng dựa vào CĐBP tỏ ra hiệu quả hơn so với ELR.

3.2 Ước lượng điểm chuyển

Tiếp theo ta sẽ so sánh hiệu quả ước lượng điểm chuyển τ cho mô hình $y = h(x-\tau)^+ + \varepsilon$ bằng phương pháp CĐBP và phương pháp SEG thông qua mô phỏng sử dụng phần mềm R.

Để tiến hành mô phỏng chúng ta xét mô hình $y = h(x-\tau)^+ + \varepsilon$, mẫu được sinh với kích thước n=80 và sai số có phân bố chuẩn hóa. Để so sánh phương pháp CĐBP với SEG, đòi hỏi các giá trị của X không được quá liền nhau, ta sẽ sinh X từ phân bố chuẩn với trung bình 0 và độ lệch chuẩn 2, điểm chuyển thực là $\tau = 2$, tức là với $x_1 \le x_2 \le ... \le x_{80}$ được sinh từ phân bố $N(0,2^2)$ và các sai số $\varepsilon_1,...,\varepsilon_{80}$ được sinh từ phân bố chuẩn hóa $N(0,1)$. Các giá trị y_i khi đó được xác định bởi $y_i = h(x_i - 2)^+ + \varepsilon_i$. Sinh ra 1000 mẫu, trong mỗi mẫu ta sẽ ước lượng điểm chuyển, hiệu quả của phương pháp được dựa vào độ sai lệch của trung bình các điểm chuyển ước lượng trong 1000 lần lặp lại mẫu với điểm chuyển thực. Để đánh giá

hiệu quả, ta sẽ xem xét ảnh hưởng của sự thay đổi hệ số góc giữa đoạn sau so với đoạn trước ở mức nhỏ, vừa phải và lớn tương ứng với $h = 0.5,\ h = 2$ và $h = 4$.

Code R để ước lượng điểm chuyển dựa vào SEG

```
seg=function(h,s);{n=80;c=numeric(s);
for (t in 1:s);{e=rnorm(n);x=sort(rnorm(n, 0, 2))
y=numeric(n);m=length(subset(x,x<=2));
for(i in 1:m);{y[i]=e[i]};for (i in (m+1):n);
{y[i]=h*(x[i]-2)+e[i]};lm=lm(y~x);
fit=SEGMENTEDed(lm, seg.Z=~x);c[t]=fit$psi[2]} ;print(mean(c))}
```

Code R để ước lượng điểm chuyển dựa vào CĐBP

```
ss=function(s);{n=80;c=numeric(s);cp=numeric(s);al1=numeric(s);
al2=numeric(s);si1=numeric(s);si2=numeric(s); bt1=numeric(s);bt2=numeric(s);for (t in 1:s);{e=rnorm(n);
x=sort(rnorm(n, 0, 2));y=numeric(n);m=length(subset(x,x<=2));
for(i in 1:m);{y[i]=e[i]};for (i in (m+1):n);
{y[i]=h*(x[i]-2)+e[i]};e=numeric(n);r=numeric(41);for (k in 20:60);
{a1= summary (lm(y[1:k]~x[1:k]))$coefficients[1];
a2= summary (lm(y[1:k]~x[1:k]))$coefficients[2];
b1= summary (lm(y[ (k+1):n]~x[(k+1):n]))$coefficients[1]
b2= summary (lm(y[(k+1):n]~x[(k+1):n]))$coefficients[2];for (j in 1:k);
{e[j]=y[j]-(b1+b2*x[j])};for (j in (k+1):n);{e[j]=y[j]-(a1+a2*x[j])};r[k-19]=
sum(e^2)};cp[t]=which.max(r)+19;
c1= summary (lm(y[1: cp[t] ]~x[1: cp[t]]))$coefficients[1];
c2= summary (lm(y[1: cp[t]]~x[1: cp[t]]))$coefficients[2];
z=numeric(n-cp[t]);for (j in 1:(n-cp[t]));
{z[j]=y[j+cp[t]]-(c1+c2*x[j+cp[t]])} ;
d1= summary (lm(z~x[(cp[t]+1):n]))$coefficients[1];
d2= summary (lm(z~x[(cp[t]+1):n]))$coefficients[2];u=-d1/d2;
if(u>x[cp[t]]&u<x[cp[t]+1]){c[t]=u};
if(u<=x[cp[t]]){c[t]=x[cp[t]]} ;if(u>=x[cp[t]+1]){c[t]=x[cp[t]+1]}};
print(mean(c))}
```

Kết quả thu được thể hiện ở bảng sau:

Bảng 5. Trung bình điểm chuyển trong 1000 lần sinh mẫu

	h=0.5		h=2		h=4	
	SEG	CĐBP	SEG	CĐBP	SEG	CĐBP
$\overline{\hat{\tau}}$		1.4943196	*1.904715*	1.619633	*1.9961*	1.676183

Nhận thấy rằng trong bảng trên, trong trường hợp cho ra kết quả (với $h = 2;\ 4$) thì phương pháp SEG tỏ ra hiệu quả hơn so với phương pháp CĐBP. Tuy nhiên, trong trường hợp độ lệch hệ số góc giữa đoạn sau với đoạn trước là nhỏ thì phương pháp SEG không cho ra kết quả (vì các giá trị quan sát của đoạn sau và đoạn trước sai lệch nhỏ)- trong trường hợp này thì phương pháp CĐBP là một giải pháp.

4. KẾT LUẬN

Bằng việc sử dụng phần mềm R trong việc mô phỏng so sánh hiệu quả ước lượng thời điểm chuyển và điểm chuyển trong mô hình hồi quy tuyến tính có gãy liên tục, chúng ta thu được: khi mô hình có sai số tuân theo quy luật chuẩn và cả không có phân bố chuẩn, cả trong trường hợp phân bố của biến giải thích là chuẩn hay đều, để ước lượng thời điểm chuyển thì phương pháp CĐBP tỏ ra hiệu quả hơn phương pháp ELR được đề xuất bởi Liu Z. và Qian L., và theo như Liu Z. và Qian L. tiến hành mô phỏng thì phương pháp ELR hiệu quả hơn so với phương pháp SEG của Muggeo, V.M.R. Trong trường hợp ước lượng điểm chuyển, phương pháp SEG tỏ ra hiệu quả hơn so với phương pháp CĐBP, tuy nhiên khi độ dao động của hệ số góc giữa đoạn sau so với đoạn trước là nhỏ thì phương pháp SEG không giải quyết được bài toán ước lượng, khi đó phương pháp CĐBP là một giải pháp - tuy chưa được tốt nhưng cho được ước lượng của điểm chuyển.

TÀI LIỆU THAM KHẢO

1. To Van Ban, Nguyen Thi Quyen (2016), "Estimatng a change – point in two-phases regression model based on the shift of parameter estimates", *Theoretical Mathematics and Applications*, 6(4), pp.33-52.
2. Owen A. (1991), "Empirical likelihood for linear model", *Ann. Statist*., 19(4), 1725-1747.
3. Muggeo, V.M.R. (2003), "Estimating regression models with unknown break-point", *Statistics in Medicine,* 22, 3055–3071.
4. Liu Z., Qian L. (2009), "Changepoint estimation in a SEGMENTEDed linear re-gression via empirical likelihood", *Communications in Statistics-Simulation and Coputation*,39, 85-100.