The Korean Communications in Statistics
Vol. 13 No. 1, 2006, pp. 125-134

Interval Regression Models Using Variable Selection[1)]

Seung Hoe Choi[2)]

Abstract

This study confirms that the regression model of endpoint of interval outputs is not identical with that of the other endpoint of interval outputs in interval regression models proposed by Tanaka et al. (1987) and constructs interval regression models using the best regression model given by variable selection. Also, this paper suggests a method to minimize the sum of lengths of a symmetric difference among observed and predicted interval outputs in order to estimate interval regression coefficients in the proposed model.

Some examples show that the interval regression model proposed in this study is more accuracy than that introduced by Inuiguchi et al. (2001).

Keywords : Interval Regression; Least Squares Method; Symmetric Difference.

1.서론

자연과학, 공학, 그리고 사회과학 등 여러 분야에서 독립변수와 종속변수 사이의 함수적인 관계를 규명하고 설명하기 위하여 수학적 모형을 가정하고, 수집된 자료를 이용하여 제안된 모형을 추정하고 분석하는 방법을 회귀분석이라 한다. 회귀분석에 사용되는 종속변수는 키와 몸무게, 수입과 지출, 일조량과 수확량 등과 같이 수로 표현되는 경우도 있지만 때로는 주식정보, 혈압, 그리고 비행기의 이륙과 착륙에 대한 정보를 제공하는 시정이나 운고와 같이 한 변수의 성질을 일정한 범위로 설명하는 종속변수는 수가 아닌 구간으로 표현되는 경우도 존재한다. 따라서 종속변수가 구간으로 표현되는 회귀모형인 구간회귀모형(interval regression model)에 대한 통계적인 성질을 연구할 필요가 있다.

Tanaka 등(1987, 1992)은 종속변수가 구간으로 표현되는 다음의 구간회귀모형

$$Y_i = A_1 x_{i1} + \cdots + A_p x_{ip} \tag{1.1}$$

과 수치해석적인 모형추정방법을 소개하였다. 구간회귀모형 (1.1)에서 x_{ij}는 양수인 독립변수이고, $Y_i = \{x : l_{y_i} \le x \le r_{y_i}\}$는 구간으로 표현된 종속변수이며, $A_j = \{x : a_j^l \le x \le a_j^r\}$는 구간회귀계수(interval regression coefficient)이다. 그리고 Y_i와 A_j는 각각 $[l_{y_i}, r_{y_i}]$와 $[a_j^l, a_j^r]$로 표시한다(Hwang과 Seo (1989)를 참조).

---

1) This research has been supported by 2005 Hankuk Aviation Faculty Research Grant.

2) Associate Professor, Department of General Studies, Hankuk Aviation University, Koyang, Kyungkido, 412-791. Korea. E-mail : shchoi@hau.ac.kr

구간회귀모형의 구간회귀계수를 추정하기 위하여 발표된 많은 논문들은 추정된 구간의 오른쪽 끝점과 왼쪽 끝점의 차인 추정된 구간의 길이를 최소로 하는 수치해석적인 방법을 사용하였다. Tanaka 등(1998, 1999)은 추정된 구간회귀모형의 예측된 구간이 관찰된 구간을 포함하는 경우인 확대모형(possibilistic model)과 관찰된 구간이 예측된 구간을 포함하는 경우인 축소모형(necessity model)을 소개하고, 제안된 각각의 구간회귀모형을 추정하기 위하여 수치해석적방법인 선형계획법과 비선형계획법을 이용하였다. 또한, 이상치를 포함하고 있는 구간회귀모형을 추정하기 위하여 회귀분위추정량(regression quantile estimator)을 이용하였다. Hong과 Hwang (2004, 2005) 그리고 Jeng 등(2003)은 Tanaka 등(1987)이 제안한 구간선형회귀모형뿐만 아니라 구간비선형회귀모형을 추정하기 위하여 Support Vector Machine(SUM)을 이용하였다. 구간회귀모형에 관련된 대부분의 논문들이 수치해석적인 방법을 사용하였으나 Inuiguchi 등(2001)은 Tanaka 등(1998)이 소개한 축소모형, 확대모형과 함께 예측된 구간과 관찰된 구간이 서로 교차하는 근사모형(approximate model)을 소개하고 구간회귀계수를 추정하기 위하여 관측된 구간과 예측된 구간사이의 거리를 최소화하는 통계적인 방법을 사용하였다. Inuiguchi 등(2001)은 두 구간사이의 거리를 정의하기 위하여 Diamond (1988)가 제안한 구간사이의 거리와 민코브스키(Minkowski) 차를 이용하였다. 지금까지 발표된 구간회귀모형에 관련된 논문은 모두 구간회귀모형의 양 끝점에 대한 회귀모형이 일치하는 것으로 가정하였다.

그러나 구간회귀모형의 양 끝점에 대한 회귀모형이 반드시 일치하지 않는 경우도 있다. 독립변수와 주어진 구간의 한쪽 끝점에 대한 회귀모형은 이차식이지만 또 다른 끝점에 대한 회귀모형은 지수식이나 일차식인 경우도 존재한다. 회귀분석에서 사용되는 이분산성처럼 구간회귀계수의 한쪽 끝점에 대한 회귀모형은 선형인지만 주어진 구간의 길이에 대한 회귀모형은 상수가 아닌 경우도 존재한다(Chen (1999)을 참조). 따라서 구간회귀계수의 양 쪽 끝점에 대한 수학적인 관계가 반드시 일치할 필요는 없으며 적합하지 않은 회귀모형을 이용함으로 잘못된 구간회귀모형을 추정할 수도 있다. 또한 독립변수와 길이에 대한 회귀모형은 (1.1)과 같이 선형으로 표현됨으로 독립변수의 값이 증가할수록 추정된 구간의 길이도 증가하는 경우도 존재한다. 추정된 구간회귀모형의 정확성은 추정된 구간의 길이에 영향을 받음으로 구간으로 추정된 모형의 정확성을 높이기 위하여 주어진 종속변수의 길이와 독립변수에 대한 적절한 반응함수를 이용할 필요가 있다.

본 논문에서는 구간회귀모형에서 주어진 구간회귀모형의 양 끝점에 대한 회귀모형이 일치하지 않는 경우가 있음을 확인하고, 회귀분석에서 사용하는 변수선택법을 이용하여 주어진 독립변수와 구간회귀모형의 끝점과 길이에 대한 최적의 회귀모형을 유도한다. 그리고 구간회귀모형의 끝점과 길이에 대한 최적의 회귀모형을 이용하여 관측된 구간과 예측된 구간에 대한 대칭차의 길이를 최소화하는 구간최소제곱법으로 구간회귀모형을 추정한다. 또한 예제를 통하여 Inuiguchi 등(2001)이 추정한 구간회귀모형과 비교하여 본 논문에서 제시된 추정법의 정확성을 조사한다.

2. 구간회귀모형

본 절에서는 구간에 대한 기본적인 연산과 특별한 연산에 대한 결과를 제시하고, 구간회귀모형의 양 끝점에 대한 회귀모형이 반드시 일치하지 않을 수 있음을 확인한다. 그리고 구간에 대한 대칭차의 길이를 최소화하는 방법으로 구간회귀모형을 추정한다.

두 구간 $I_1 = [l_1, r_1]$와 $I_2 = [l_2, r_2]$에 대한 연산의 정의, 민코브스키 차(Minkoski difference, $\ominus_M$) 그리고 대칭차(symmetric difference, $\triangle$)는 다음과 같다. 여기서 I^C는 구간 I의 여집합(complement set)이고, k는 실수이며 α와 β는 각각 집합 $\{l_1 l_2, l_1 r_2, r_1 l_2, r_1 r_2\}$의 최소값과 최대값이다.

$$(p_1)\ [l_1, r_1] \pm k = [l_1 \pm k, r_1 \pm k]\ ,$$

$$(p_2)\ k \cdot [l_1, r_1] = \begin{cases} [kl_1, kr_1] & \text{if } k \geq 0 \\ [kr_1, kl_1] & \text{if } k < 0 \end{cases}\ ,$$

$$(p_3)\ I_1 \ominus_M I_2 = \bigcap s(I_1 - s) \text{ for } s \in I_2\ ,$$

$$(p_4)\ I_1 \triangle l_2 = (I_1^C \bigcap I_2) \bigcup (I_1 \bigcap I_2^C)\ ,$$

$$(p_5)\ I_1 \cdot I_2 = [\alpha, \beta]\ .$$

이제 구간회귀모형의 양쪽 끝점에 대한 회귀모형에 대하여 알아보자. 구간에 대한 기본적인 연산의 결과로부터 Tanaka 등(1987)이 제시한 구간회귀모형 (1.1)의 양쪽 끝점에 대한 회귀모형은 일치함을 알 수 있다. 다음 예제는 주어진 구간회귀모형의 양쪽 끝점에 대한 회귀모형이 일치하지 않음을 보여준다.

예제 1. Tanaka와 Lee (1998) 그리고 Hong와 Hwang (2005)은 다음 표의 자료를 이용하여 구간회귀모형 (1.1)에 대한 확장모형과 축소모형을 추정하였다.

<표 1> 구간회귀모형에 대한 Tanaka의 자료

x	1	2	3	4	5	6	7	8
Y	[15,30]	[20,37.5]	[15,35]	[25,60]	[25,55]	[40,65]	[55,95]	[70,100]

관측된 구간의 왼쪽 끝점(l)과 오른쪽 끝점(r)에 대한 최적의 회귀모형을 찾기 위하여 회귀분석에서 사용하는 변수선택법을 x와 x^2에 적용한 결과, 독립변수에 의해서 설명되는 종속변수의 총 변동의 비율을 표시하는 수정된 R^2의 값을 <표 2>와 같이 얻었다.

<표 2>는 구간회귀모형의 왼쪽 끝점은 변수 x와 x^2을 사용하여 설명하고, 오른쪽 끝점을 설명하기 위해서는 x^2을 사용하는 것이 타당함을 보여주고 있다. 따라서 구간회귀모형의 양 끝점에 대한 최적의 회귀모형은 일치하지 않음을 확인할 수 있다.

<표 2> Tanaka의 자료에 대한 변수선택법의 결과

변수		x	x^2	(x, x^2)
수정된 R^2	l	0.8169	0.9454	0.9707
	r	0.8864	0.925	0.9114

예제2. 아래 표에 제시된 자료는 Diamond (1988)가 퍼지회귀모형을 추정하기 위하여 소개한 것으로 독립변수(x)는 수이고 종속변수(Y)는 구간인 구간회귀모형에 대한 자료이다.

<표 3> 구간회귀모형에 대한 Diamond의 자료

x	21	15	15	9	12	18	6	12
Y	[3.2, 4.8]	[2.7,3.3]	[3.15,3.85]	[1.6,2.4]	[2.55,3.45]	[2.8,4.2]	[2.12,2.88]	[2,3]

<표 3>에서 주어진 구간의 왼쪽 끝점(l)과 오른쪽 끝점(r)을 가장 잘 설명하는 회귀모형을 찾기 위하여 변수선택법을 x와 x^2에 적용한 결과, 수정된 R^2와 Mallow c_p의 값은 <표 4>와 같다.

<표 4> Diamond의 자료에 대한 변수선택법의 결과

변수		x	x^2	(x, x^2)
l	수정된 R^2	0.5982	0.5892	0.5208
	c_p	1.0312	1.1434	3.000
r	수정된 R^2	0.7871	0.851	0.8308
	c_p	3.5508	1.2842	3.000

<표 4>에 주어진 수정된 R^2와 Mallow c_p의 값은 구간회귀모형의 왼쪽 끝점과 오른쪽 끝점에 대한 적절한 설명변수는 각각 x와 x^2임을 설명한다. 따라서 <표 3>에서 제시된 자료의 양 끝점에 대한 회귀모형은 동일하지 않음을 확인할 수 있다.

이제 구간회귀모형에서 주어진 구간회귀계수를 추정하는 방법에 대하여 생각하여 보자. 회귀분석에서 일반적으로 사용되는 최소제곱법을 구간회귀모형에서 동일하게 적용할 수 없다. 그 이유는 두 구간에 대한 차의 길이는 각 구간 길이의 합으로 표현되기 때문이다. 즉, 동일한 구간의 차에 대한 길이는 주어진 구간 길이의 두 배가 된다. 따라서 관측된 구간과 예측된 구간의 차를 최소화하는 방법을 구간회귀분석에서는 이용할 수 없음으로 다른 방법을 생각하여야 한다.

구간에 대한 차의 대안으로 두 집합에 대한 차를 의미하는 집합의 대칭차를 생각하자. 두 구간의 교집합이 공집합이며 두 구간에 대한 대칭차는 두 구간의 합집합이고,

두 구간의 교집합이 공집합이 아니며 주어진 구간에 대한 대칭차는 두 구간의 합집합에서 교집합을 뺀 집합이다. 따라서 구간회귀모형에서 주어진 구간오차의 길이를 최소화하기 위해 두 구간에 대한 대칭차를 이용할 수 있다.

주어진 구간에 대한 양 끝점의 차이로 정의된 구간 $I_i=[l_i,r_i]$의 길이를 $m(I_i)$로 표시하자. 그러면 $m(I_i)=r_i-l_i$이고 주어진 두 구간 I_1과 I_2에 대한 대칭차의 길이는 다음과 같다.

$$m(I_1 \triangle I_1)=0\,,$$

$$m(I_1 \triangle I_2)=\begin{cases}|l_1-l_2|+|r_1-r_2|, & \text{if } I_1 \bigcap I_2 \neq \phi\,,\\ (r_2+r_1)-(l_1+l_2), & \text{if } I_1 \bigcap I_2 = \phi\,.\end{cases}$$

또한 구간 연산으로부터 구간 I_i는 $l_i+[0,r_i-l_i]$과 같이 표현할 수 있으므로 구간회귀모형 (1.1)은

$$Y_i=\sum_{j=1}^{p} a_j^l x_{ij}+\sum_{j=1}^{p}[0,(a_j^r-a_j^l)x_{ij}] \tag{2.1}$$

와 같이 표현할 수 있다. 즉, 구간회귀모형은 구간회귀계수의 왼쪽 끝점에 대한 회귀모형과 구간회귀계수가 $[0,(a_j^r-a_j^l)\,x_{ij}]$인 구간회귀모형으로 구분하여 표현할 수 있다. 식 (2.1)을 이용하여 종속변수가 구간으로 표현되는 구간회귀모형을 수와 구간에 대한 회귀모형으로 일반화할 수 있다.

본 논문에서는 다음과 같은 구간회귀모형을 생각한다.

$$Y_i=f(C,X_i)+g(I,X_i)+E_i$$

(2.2)모형 (2.2)에서 함수 f와 g는 실가함수이고, $E_i=[l_{e_i},r_{e_i}]$는 구간오차이며 $C=(c_1,\cdots,c_p)$와 $I=(I_1,\cdots,I_p)$에서 c_j는 실수이고 I_j는 왼쪽 끝점은 0이고 오른쪽 끝점인 s_j인 구간 $[0,s_j]$을 표시한다. 모형 (2.2)에서 함수 f와 g가 동일한 선형함수이면 모형 (2.2)는 모형 (1.1)과 동일함을 구간 연산의 성질에 의하여 쉽게 확인할 수 있다. 즉, 모형 (2.2)는 모형 (1.1)의 확장을 의미한다. 그리고 독립변수 x_{i1}이 1이고, 구간회귀계수의 길이인 s_1을 제외한 모든 $s_j(2 \le j \le p)$가 0이면 모형 (2.2)는 구간회귀계수의 왼쪽 끝점과 독립변수는 수학적인 함수관계가 있으나 구간회귀계수의 길이는 독립변수에 영향을 받지 않는 구간회귀모형이 된다. 즉, 구간회귀계수의 왼쪽 끝점은 독립변수의 크기에 따라 영향을 받지만 구간회귀계수의 길이는 항상 일정한 구간회귀모형이 된다.

구간회귀모형 (2.2)에 포함된 구간회귀계수 c_j와 I_j를 추정하기 위해 오차구간을

$$\begin{aligned}E_i &= l_{e_i}+[0,s_{e_i}]\\ &= Y_i-f(C,X_i)-g(I,X_i)\\ &= l_{y_i}-f(C,X_i)+[0,s_{y_i}]-g(I,X_i)\end{aligned}$$

와 같이 변형할 수 있다. 여기서 $s_{e_i}=r_{e_i}-l_{e_i}$이고 $s_{y_i}=r_{y_i}-l_{y_i}$이다. 위 식을 이용하여 구간회귀모형을 다음과 같이 구간과 수로 구분된 회귀모형으로 표현할 수 있다.

$$(r_1)\ \ l_{y_i} = f(C, X_i) + l_{e_i}\ ,$$
$$(r_2)\ \ [0, s_{y_i}] = g(I, X_i) + [0, s_{e_i}]\ .$$

모형 r_1에 포함된 회귀계수 c_j는 회귀분석에서 사용하는 방법으로 추정할 수 있고 구간으로 표현된 회귀모형 r_2의 구간회귀계수를 추정하기 위해서 구간에 대한 대칭차의 길이를 이용한다. 구간으로 표현된 $g(I, X_i)$의 왼쪽 끝점은 0이고 오른쪽 끝점은 $g(S, X_i)$이므로 모형 r_2에서 주어진 두 구간 $[0, s_{y_i}]$와 $g(I, X_i)$에 대한 대칭차의 길이는 두 구간의 오른쪽 끝점 차인 $|s_{y_i} - g(S, X_i)|$와 같고 $S = (s_1, \cdots, s_p)$이다. 따라서 모형 r_2의 구간회귀계수는 두 구간에 대한 대칭차의 길이인 $|s_{y_i} - g(S, X_i)|$의 총합을 최소화하는 방법으로 추정할 수 있다.

회귀분석에서 사용하는 변수선택법을 이용하여 독립변수 X_i와 구간으로 주어진 구간자료 Y_i의 왼쪽 끝점(l_{y_i})과 길이(s_{y_i})에 대한 최적의 회귀모형을 구한 후 다음과 같은 함수

$$\sum_{i=1}^{n} \rho(l_{y_i} - f(x_{ij}, c_j))$$

와

$$\sum_{i=1}^{n} \psi(s_{y_i} - g(x_{ij}, s_j))$$

을 최소로 하는 값을 구간 M-추정량(interval M-estimator)이라 하자. 여기서 s_j는 0보다 작지 않은 수이며, 함수 ρ와 ψ는 R에서 R^+로 가는 연속함수로서 0에서 유일한 최소값을 갖는다. 만약 $\rho(x) = \psi(x) = x^2$이면 구간최소제곱추정량(interval least square estimator)이라 하고 $\widehat{Y}_i$와 같이 표시하자. 그리고 $\rho(x) = \psi(x) = |x|$이면 구간최소절대편차추정량(interval least deviation estimator)이라 하고 $\widetilde{Y}_i$와 같이 표시하자.

다음 절에서는 본 절에서 제시된 추정법을 사용하여 구간회귀모형을 추정하고, 추정된 구간회귀모형의 정확성을 Inuiguchi 등(2001)이 제시한 모형과 비교하여 설명한다.

3. 구간회귀추정량의 정확성

본 절에서는 앞 절에서 소개된 구간회귀계수의 왼쪽 끝점과 길이에 대한 최적의 회귀모형을 이용하여 표현된 구간회귀모형을 추정하기 위하여 구간최소제곱추정량을 이용한다. 또한 본 논문에서 제시된 추정법의 정확성을 조사하기 위하여 Inuiguchi 등(2001)이 추정한 구간회귀모형과 비교한다.

구간회귀모형의 추정법에는 추정된 구간에 대한 길이의 총합을 최소화 하는 수치해석적인 방법과 관측된 구간과 예측된 구간에 대한 차의 총합을 최소로 하는 통계적인 방법이 있다. Inuiguchi 등(2001)은 구간회귀계수에 대한 최적모형을 고려하지 않

은 구간회귀모형(1.1)에 포함된 구간회귀계수를 추정하기 위하여 민코브스키 차, 다이아몬드 거리, 구간의 곱 그리고 실수의 절대값과 비슷한 개념인 절대구간을 이용하였다. 수치해석적 방법으로 추정된 구간의 길이의 총합을 비교하여 추정된 회귀모형의 정확성을 비교할 수 있지만 잘 정의된 두 구간에 대한 차의 총합을 최소로 하는 방법을 이용하여 추정된 구간회귀모형의 정확성을 조사할 수 있다.

앞 절에서 언급한 것처럼 두 구간의 차에 대한 길이는 각 구간에 대한 길이의 합으로 정의됨으로 두 구간에 대한 차를 새롭게 정의하여야 한다. 구간회귀모형을 추정하기 위하여 제안된 추정량의 정확성을 비교하기 위하여 다음과 같은

$$C(Y,\hat{Y}) = \sum_{i\in A} m(Y_i \Delta \hat{Y}_i) + \sum_{i\notin A} h(Y_i, \hat{Y}_i) \tag{3.1}$$

척도를 생각하자. 여기서 $A = \{i : Y_i \cap \hat{Y}_i \neq \phi\}$이고

$$h(Y_i, \hat{Y}_i) = \min\{|l_{y_i} - \hat{r}_{y_i}|, |r_{y_i} - \hat{l}_{y_i}|\} + (r_{y_i} - l_{y_i}) + (\hat{r}_{y_i} - \hat{l}_{y_i})$$

이다. 관측된 구간과 예측된 구간의 모든 교집합이 공집합이 아니면 식 (3.1)에서 주어진 A는 공집합이 된다. 즉, 관측된 구간과 예측된 구간의 각 끝점에 대한 차의 총합을 나타낸다. 관찰된 구간과 예측된 구간이 서로 가까이 있으면 있을수록 척도 $C(Y,\hat{Y})$는 점점 작아짐으로 척도 (3.1)를 이용하여 추정된 구간회귀모형의 정확성을 비교할 수 있다. 즉, 구간사이의 오차를 나타내는 $C(Y,\hat{Y})$의 값이 작을수록 추정된 구간회귀모형의 정확성은 좋다고 생각할 수 있다.

예제3. 아래 자료는 Tanaka 등(1987)이 선형계획법과 비선형계획법을 이용하여 구간회귀모형을 추정하기 위해 소개한 자료의 왼쪽 끝점(l)과 구간의 길이(s)이다. 동일한 자료를 이용하여 Inuiguchi 등(2001)은 민코브스키 차와 다이아몬드 거리를 이용하여 구간회귀모형을 추정하였다.

<표 5> 구간회귀모형에 대한 Tanaka의 자료

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
l	0.19	0.24	0.20	0.20	0.22	0.22	0.35	0.37	0.41
s	0.10	0.08	0.07	0.26	0.16	0.11	0.21	0.23	0.48

<표 6>은 구간회귀모형을 추정하기 위하여 관측된 구간의 왼쪽 끝점과 구간의 길이에 대한 변수선택법을 이용한 결과인 수정된 R^2와 Mallow c_p의 값이다.

<표 6>은 <표 5>에서 제시된 자료에 대한 구간회귀모형의 왼쪽 끝점은 변수 x와 x^2을, 그리고 구간회귀모형의 길이는 변수 x^2을 사용하는 것이 타당함을 보여주고 있다. 따라서 변수선택법를 이용한 구간회귀모형은

$$f(C, X_i) + g(I, X_i) = c_0 + c_1 x_i + c_2 x_i^2 + [0, s_0] + [0, s_1] \cdot x_i^2$$

와 같다.

<표 6> Tanaka의 자료에 대한 변수선택법의 결과

변수		x	x^2	(x, x^2)
l	수정된 R^2	0.6874	0.8089	0.8535
	c_p	9.9407	4.1311	3.0000
s	수정된 R^2	0.4892	0.5801	0.5656
	c_p	3.2311	1.7667	3.0000

예제 4. <표 5>에서 제시된 자료를 이용하여 구간최소제곱추정량과 Inuiguchi 등(2001)의 방법으로 추정한 구간회귀모형을 비교하여 보자. <표 5>에서 제시된 자료에 대한 구간회귀모형을 추정하기 위하여 다이아몬드 거리를 이용한 Inuiguchi 등(2001)의 결과는

$$\widehat{Y}_i = [\widehat{l_{y_i}}, \widehat{r_{y_i}}] = [0.30789, 0.42411] - 0.132x + [0.024407, 0.04893]x^2$$

이고 민코브스키 차를 이용한 결과는

$$\breve{Y}_i = [\breve{y_i^c}, \breve{y_i^r}] = [0.266272, 0.383728] - 0.118333x + [0.027213, 0.046121]x^2$$

와 같다. Inuiguchi 등(2001)이 구간회귀모형을 설명하기 위하여 사용한 자료를 본 논문에서 제안한 구간최소제곱추정량을 이용하여 추정한 구간회귀모형은 다음과 같다.

$$\widehat{Y}_i = [\widehat{l_{y_i}}, \widehat{r_{y_i}}] = [0.27052, 0.32863] - 0.08025x + [0.02221, 0.03447]x^2 .$$

주어진 구간의 왼쪽 끝점과 길이에 대한 최적의 회귀모형을 이용하여 추정한 구간회귀모형과 Inuiguchi 등(2001)이 추정한 회귀모형은 모두 일차식의 구간회귀계수는 구간이 아닌 수로 추정하였다. 그러나 추정된 구간회귀모형의 x^2에 대한 구간회귀계수의 길이는 차이가 있었다.

<표 7>은 Inuiguchi 등(2001)의 방법과 구간최소추정량을 이용하여 추정된 구간의 길이를 보여준다. <표 7>로부터 구간최소제곱추정량을 이용하여 추정된 구간에 대한 길이의 총합($\hat{s}_{y_i}$) 3.8667는 다이아몬드 거리와 민코브스키 차를 이용하여 추정된 구간에 대한 길이의 총합 6.9642($\breve{s}_{y_i}$)와 6.9642($\breve{s}_{y_i}$)보다 작음을 확인할 수 있다. 추정된 구간에 대한 길이의 총합을 최소화하는 수치해석적인 방법에서도 구간최소추정량을 이용하여 추정한 구간회귀모형이 더 정확할 수 있음을 보여준다.

추정된 구간과 관측된 구간에 대한 대칭차의 길이로 정의된 척도 (3.1)를 이용하여 구간최소제곱추정량과 Inuiguchi 등(2001)의 방법에 대한 정확도를 비교한 결과를 <표 8>에서 확인할 수 있다. <표 8>은 Inuiguchi 등(2001)이 다이아몬드 거리와 민코브스키 차를 사용하여 추정한 구간과 관측된 구간에 대한 대칭차의 길이 총합이 각각 5.2846와 4.36718이고 구간최소제곱추정량을 이용하여 추정한 구간과 주어진 구간에 대한 대칭차의 길이 총합이 2.36086임을 보여준다.Inuiguchi 등(2001)이 이용한 <표 5>의 자료에서는 구간최소제곱추정량을 이용하여 추정한 구간회귀모형이 민코브스키 차와 다이아몬드 거리를 이용하여 추정한 Inuiguchi 등(2001)의 구간회귀모형보다 더 정확할 수 있음을 보여준다.

<표 7> 추정된 구간의 길이

x_i	구간 길이		
	$\hat{s}_{y_i}$	$\check{s}_{y_i}$	$\hat{s}_{y_i}$
1	0.27274	0.25469	0.15062
1.5	0.36940	0.33749	0.20607
2	0.47831	0.42975	0.26765
2.5	0.59949	0.53146	0.33536
3	0.73293	0.64262	0.40920
3.5	0.87863	0.76323	0.48917
4	1.03659	0.89330	0.57527
4.5	1.20681	1.03283	0.66750
5	1.38930	1.18181	0.76586
Total	6.9642	6.06718	3.86670

<표 8> 각 추정량에 대칭차의 길이

x_i	대칭차 길이		
	$m(Y_i \triangle \hat{Y}_i)$	$m(Y_i \triangle \check{Y}_i)$	$m(Y_i \triangle \hat{Y}_i)$
1	0.19334	0.15469	0.09558
1.5	0.28940	0.25749	0.12607
2	0.40831	0.35975	0.19765
2.5	0.33949	0.27146	0.09277
3	0.57293	0.48262	0.26852
3.5	0.76863	0.65323	0.46260
4	0.82659	0.68330	0.36527
4.5	0.97681	0.80283	0.43750
5	0.90930	0.70181	0.31490
Total	5.2848	4.36718	2.36086

비록 예제를 통하여 회귀분석에서 사용하는 변수선택법과 구간최소제곱추정량을 이용한 구간회귀모형이 Inuiguchi 등(2001)이 제시한 방법을 이용하여 추정한 구간회귀모형보다 더 정확할 수 있음을 설명하였으나 이 결과를 일반적인 구간회귀모형으로 확장할 수는 없다. 일반적인 구간회귀모형에 대한 정확성을 판단하기 위해서는 회귀분석처럼 구간회귀모형에 대한 통계적 성질을 연구하여야 한다.

4. 결론

본 논문에서는 Tanaka 등(1987)이 소개한 구간회귀모형을 추정하는 방법에 대하여 연구하였다. 주어진 구간회귀모형의 양 쪽 끝점에 대한 회귀모형이 반드시 일치하지

않음을 확인하고, 구간회귀모형을 추정하기 위하여 회귀분석에서 사용하는 변수선택법을 사용하여 유도한 구간회귀모형의 끝점과 길이에 대한 최적의 회귀모형을 이용하였다. 구간회귀모형에 포함된 구간회귀계수를 추정하기 위하여 관측된 구간과 예측된 구간 사이의 대칭차에 대한 길이를 최소화하는 방법인 구간최소추정량을 이용하였다. 예제를 통하여 본 논문에서 제시된 추정법이 민코브스키 차와 다이아몬드 길이를 이용한 Inuiguchi 등(2001)의 추정법보다 더 정확할 수 있음을 확인하였다.

구간회귀모형에 대한 정확성을 일반적으로 비교하기 위하여 구간회귀모형을 추정하기 위해 사용되는 각 추정량에 대한 통계적 성질을 계속 연구할 필요가 있다.

참고문헌

[1] Hwang, S.G. and Seo, Y.J. (1989). 제약부 구간 선형 회귀모델에 의한 실동시간의 견적. *Journal of the Korean OR/MS Society*, Vol. 14, 105-114.

[2] Chen, Y. (1999), Fuzzy ranking and quadratic fuzzy regression. *Computers and Mathematics with Applications*, Vol. 38, 265-279.

[3] Diamond, P. (1988), Fuzzy least squares. *Information Science*, Vol. 46, 141-157.

[4] Hong, D.H. and Hwang, C. (2004). Support vector Machine for Internal Regression. *Proceeding of Autumn Conference on Korean Statistical Society*, 67-72.

[5] Hong, D.H. and Hwang, C. (2005). Internal regression analysis using quadratic loss support vector machine. *IEEE Transactions on Fuzzy Systems*, Vol. 13, 229-237.

[6] Inuiguchi, M., Fujita, H. and Tanino, T. (2001). Interval linear regression analysis bases on Minkowski difference, Proceeding of International Conference on Information Systems. *Analysis and Synthesis*, Vol. 7, 112-117.

[7] Ishibuchi, H. and Tanaka, H. (1992). Fuzzy regression analysis using neural networks. *Fuzzy sets and Systems*, Vol. 50, 57-65.

[8] Jeng, J.T., Chuang, C. and Su, S.F. (2003). Support vector interval regression networks for interval regression analysis. *Fuzzy Sets and Systems*, Vol. 138, 283-300.

[9] Lee, H. and Tanaka, H. (1999). Upper and lower approximation models in interval regression using regression quantile techniques. *European Journal of Operational Research*, Vol. 116, 653-666.

[10] Tanaka, H., Hayashi, I. and Watada, J. (1987). Interval regression analysis. *Third Fuzzy System Symposium*, 9-12.

[11] Tanaka, H. and Lee, H. (1998). Interval regression analysis by quadratic programming approach. *IEEE Transactions on Fuzzy Systems*, Vol. 6, 473-481.

[Received October 2005, Accepted February 2006]