한 국 통 계 학 회 논 문 집
2009, 16권, 5호, 781–790

분포무관추정량을 이용한 퍼지회귀모형

윤진희[a], 최승회[1,b]

[a]연세대학교 경제학과, [b]한국항공대학교 인문자연학부

요 약

본 논문에서는 퍼지수를 포함한 모수적 회귀모형을 추정하기 위하여 분포무관추정량으로 알려진 순위변환방법과 Theil 방법을 소개한다. 순위변환방법은 퍼지수의 α-수준집합의 중심과 폭에 대한 순위를 이용하고 Theil 방법은 α-수준집합의 중심과 폭에 대한 추정한 값들의 중위수를 이용한다. 예제를 이용하여 분포무관추정량으로 추정된 퍼지회귀모형의 효율성을 최소자승법과 여러 가지 방법으로 추정된 퍼지회귀모형과 비교한다.

주요용어: 퍼지회귀모형, 순위변환방법, Theil 방법, α-수준집합.

1. 서론

Tanaka 등 (1982)은 애매하거나 불확실하게 표현된 변수사이의 인과관계를 설명하기 위해 Zadeh (1965)가 소개한 퍼지집합을 회귀모형에 적용한 퍼지회귀모형을 소개하였다. 퍼지회귀모형은 설명변수에 따라 변하는 종속변수의 함수, 즉 반응함수의 모양에 따라 두 가지로 구분할 수 있다. 반응변수와 독립변수간의 함수인 회귀방정식을 모르는 경우와 두 변수 사이의 관계를 아는 경우로 구분하여 전자와 후자를 각각 모수적 퍼지회귀모형과 비모수적 퍼지회귀모형이라 한다.

많은 사람들이 퍼지회귀모형을 추정하기 위하여 여러 가지 방법을 제시하였다. Kim과 Chen (1997)은 퍼지회귀모형과 비모수회귀모형에 대한 차이를 비교하였다. 비모수적 퍼지회귀모형을 추정하기 위하여 Wang 등 (2007)은 커널방법과 국부선형방법을 제의하였다. 모수적 퍼지회귀모형을 추정하기 위해 관측퍼지와 예측퍼지수의 차의 제곱을 최소화하는 통계적인 방법과 예측퍼지수의 폭을 최소화하는 수치해석인 방법이 소개되었다.

수치해석적인 방법은 추정된 종속변수의 폭의 합을 최소화하는 방법으로 퍼지회귀모형을 추청한다. 많은 사람들이 선형 혹은 비선형 프로그래밍 방법을 이용하여 퍼지회귀모형을 연구하였다 (Kao와 Chyu, 2003; Kao와 Lin, 2005; Nasrabadi와 Nasrabadi, 2004). 추정된 퍼지수와 관측된 퍼지수의 차를 최소화하는 방법으로 퍼지회귀모형을 추정하는 통계적인 방법에 대한 결과는 많이 소개되었다 (Diamond, 1988; Choi, 2007; Kim 등, 2008).

통계적인 방법을 이용한 퍼지회귀모형에서는 주로 최소자승법이 사용되었다. 회귀분석에서최소자승추정법은 추정된 회귀모형에 대한 잔차들의 분포가 정규분포인 경우에는 좋은 추정량이지만 잔차들의 분포가 특정한 조건을 만족하지 못하면 추정된 회귀모형의 효율성은 다른 추정량보다 좋지 않을 수 있다. 이것은 이상치에 민감한 최소자승법을 이용한 퍼지회귀모형의 정확성은 떨어질 수 있음을 보여준다. 또한 퍼지회귀모형의 오차는 회귀방정식의 부정확성에서 발생한다. 따라서 추정된 퍼지회귀

[1] 교신저자: (412-791) 경기도 고양시 덕양구 화전동 200-1, 한국항공대학교 인문자연학부, 교수.
E-mail: shchoi@kau.ac.kr

모형의 효율성을 높이기 위해서 잔차들의 분포에 의존하지 않는 분포 무관 추정량을 이용할 필요가 있다.

본 논문에서는 모수적 퍼지회귀모형을 추정하기 위하여 퍼지수의 α-수준집합의 중심과 폭에 대한 순위를 이용한 순위변환방법과 관측된 α-수준집합을 이용하여 추정한 모수들의 중위수를 이용한 Theil 방법을 소개한다. 제안된 방법으로 추정된 퍼지회귀모형의 효율성을 최소자승법을 이용한 추정된 퍼지회귀모형과 비교한다.

2. 퍼지회귀모형

애매하거나 불확실하게 표현된 변수사이의 인과관계를 설명하기 위해 Tanaka 등 (1980, 1982)은 퍼지회귀모형

$$Y(X_i) = A_0 + A_1 X_{i1} + \cdots + A_p X_{ip}, \quad i = 1, \ldots, n \tag{2.1}$$

을 처음으로 소개하였다. 여기서 i번째 설명변수의 j번째 성분 X_{ij}, 회귀계수 A_j 그리고 반응변수 $Y(X_i)$는 LR-퍼지수이다. 퍼지회귀분석의 목적중 하나는 관측된 자료 $\{X_{i1}, \ldots, X_{ip}, Y_i\}$를 이용하여 관측퍼지수와 예측퍼지수 사이의 차를 가장 작게하는 회귀계수를 추정하는 것이다.

LR-퍼지수 $A = (a, l_a, r_a)_{LR}$의 소속함수(membership function)는

$$\mu_A(x) = \begin{cases} L_A\left(\dfrac{a-x}{l_a}\right), & \text{for } 0 \le a - x \le l_a, \\ R_A\left(\dfrac{x-a}{r_a}\right), & \text{for } 0 \le x - a \le r_a, \\ 0, & \text{for otherwise} \end{cases}$$

이고, 함수 L_A와 R_A는 단조증가이고

$$L_A(0) = R_A(0) = 1\text{과 } L_A(1) = R_A(1) = 0$$

을 만족한다. 여기서 a를 퍼지수 A의 중심(mode)이라 하고, l_a와 r_a를 각각 퍼지수 A의 왼쪽과 오른쪽의 폭(spread)이라 한다. 만약 $L_A(x) = R_A(x) = 1 - x$이면 LR-퍼지수 A를 삼각퍼지수라고 부르고 $(a, l_a, r_a)_T$와 같이 표시하고, 특히 양 쪽 폭이 같으면 $(a, a_s)_T$와 같이 표시한다. LR-퍼지수 $A = (a, l_a, r_a)_{LR}$의 α-수준집합은

$$A(\alpha) = \begin{cases} \{x|\mu_A(x) \ge \alpha\}, & \text{if } 0 < \alpha \le 1, \\ \overline{\{x|\mu_A(x) > 0\}}, & \text{if } \alpha = 0 \end{cases}$$

이고 집합 $\overline{A}$는 A의 폐포(closure)이다. 퍼지수 A의 α-수준집합 $A(\alpha)$은 중심이 a이고, 왼쪽과 오른쪽 폭이 각각 $l_a L_A^{-1}(\alpha)$와 $r_a R_A^{-1}(\alpha)$인 폐구간이므로

$$A(\alpha) \cong \left(a,\ l_a L_A^{-1}(\alpha),\ r_a R_A^{-1}(\alpha)\right)$$

와 같이 표현할 수 있다. 따라서 관측퍼지수 $Y_i = (y_i, l_{y_i}, r_{y_i})_{LR}$의 α-수준집합은

$$\left(y_i,\ l_{y_i}(\alpha),\ r_{y_i}(\alpha)\right)$$

와 같고 $l_{y_i}(\alpha) = l_{y_i}L_{Y_i}^{-1}(\alpha)$이고 $r_{y_i}(\alpha) = r_{y_i}R_{Y_i}^{-1}(\alpha)$이다. 설명변수가 $x_{ik}(k = 1,\ldots,p)$가 실수이고 회귀계수 A_k가 퍼지수인 경우 예측퍼지수 $Y(X_i)$의 α-수준집합은

$$\left(\sum_{k=0}^{p} a_k x_{ik},\ \sum_{k=0}^{p} l_k(\alpha)x_{ik},\ \sum_{k=0}^{p} r_k(\alpha)x_{ik}\right) \tag{2.2}$$

와 같고 $l_k(\alpha) = l_k L_{A_i}^{-1}(\alpha)$이고 $r_k(\alpha) = r_k R_{A_i}^{-1}(\alpha)$이다. 그리고 설명변수 X_{ik}가 퍼지수인 경우 예측퍼지수 $Y(X_i)$의 α-수준집합은

$$\left(\sum_{k=0}^{p} a_k x_{ik},\ \sum_{k=0}^{p} l_k(\alpha)l_{X_{ik}}(\alpha),\ \sum_{k=0}^{p} r_k(\alpha)r_{X_{ik}}(\alpha)\right) \tag{2.3}$$

와 같이 표현할 수 있다. 여기서 $l_{X_{ik}}(\alpha) = l_{X_{ik}}L_{X_{ik}}^{-1}(\alpha)$이고 $r_{X_{ik}}(\alpha) = r_{X_{ik}}R_{X_{ik}}^{-1}(\alpha)$이다. 만약 식 (2.2)의 회귀계수가 $a_k = l_k(\alpha) = r_k(\alpha)$인 경우 퍼지회귀모형 (2.1)은

$$Y(X_i) = a_0 + a_1 X_{i1} + \cdots + a_p X_{ip}$$

와 같이 표현할 수 있다. 이와 같이 회귀계수가 실수인 퍼지회귀모형을 Diamond (1988), Kao와 Chyu (2003) 그리고 Kim 등 (2008)이 연구하였다. 또한 $r_{X_{ik}}(\alpha) = l_{X_{ik}}(\alpha) = x_{ik}$인 경우 식 (2.3)은 식 (2.2)와 일치한다. 따라서 본 논문에서는 식 (2.3)을 이용하여 두 변수사이의 퍼지회귀방정식을 추정한다. 퍼지회귀계수의 소속함수가 미지인 경우에는 여러 개의 α-수준집합을 추정함으로 퍼지모형을 추정하지만 회귀계수의 소속함수가 특별한 모형으로 가정할 수 있는 경우는 특정한 α-수준집합 예측함으로 퍼지회귀모형을 추정할 수 있다. 본 논문에서는 추정된 퍼지회귀모형의 효율성을 높게 하기위해 오차의 분포에 크게 영향을 받지 않는 순위변환방법과 Theil 방법을 이용하여 퍼지회귀모형의 α-수준집합을 추정한다.

2.1. 순위변환방법을 이용한 퍼지회귀모형

Iman과 Conover (1979)는 실험분석법에서 효율성이 확인된 순위변환을 회귀분석에 사용하였다. 그들은 단조 증가 자료를 포함하는 회귀모형에서 순위변환법이 전체 자료로부터 멀리 떨어진 고립된 자료에 민감하지 않는 강인추정량임을 보였다. 관측된 값과 예측된 값의 차이가 측정오차보다는 회귀모형의 부정확성에 더 의존하는 퍼지회귀모형은 일반 회귀모형보다 더 많은 고립된 자료가 존재할 가능성이 있다. 따라서 이상점에 민감한 평균을 일반화한 최소자승법보다는 순위변환방법이 더 효율적일 수 있다.

본 절에서는 반응변수의 α-수준집합의 각 성분 y_i, $l_{y_i}(\alpha)$ 그리고 $r_{y_i}(\alpha)$에 대한 추정값을 유도하기 위하여 예측퍼지수와 관측퍼지수의 각 성분에 대한 순위를 이용하는 순위변환방법을 사용한다. 관측된 설명변수 $\{(x_{i1},\ldots,x_{ip}) : i = 1,\ldots,n\}$에 대한 반응변수 y_i에 대한 예측값 $\hat{y}_i$을 추정하는 순위변환방법은 다음과 같다.

1. 관측값 $\{x_{i1},\ldots,x_{ip}\}$에 대한 $x_{ik}(k = 1,\ldots,p)$의 순위 $R(x_{ik})$와 $\{y_1,\ldots,y_n\}$에 대한 y_i의 순위 $R(y_i)$을 결정한다. 이 때 서로 다른 값의 순위가 일치할 경우에는 평균값을 순위로 결정한다.

2. 관측된 값의 순위로 구성된 집합 $\{(R(x_{i1}),\ldots,R(x_{ik}),R(y_i)) : i = 1,\ldots,n\}$에 대한 최적의 회귀모형

$$R(y_i) = \beta_0 + \sum_{k=1}^{p}\beta_k R(x_{ik}) + \sum_{j=1}^{p}\sum_{k=1}^{p}\beta_{kj}R(x_{ik})R(x_{ij}) \tag{2.4}$$

을 유도하기 위해 회귀분석에서 사용하는 변수선택법을 사용한다. 특히, 설명변수가 한 개인 단순회귀모형에서는

$$R(y_i) = \frac{(n+1)}{2} + \beta_1\left(R(x_i) - \frac{(n+1)}{2}\right)$$

을 사용한다.

3. 모형 (2.4)로부터 유도된 최적의 회귀모형과 설명변수 $\{x_{i1}, \ldots, x_{ip}\}$에 대한 순위집합 $\{R(x_{i1}), \ldots, R(x_{ip})\}$을 이용하여 반응변수의 중심의 순위 $R(y_i)$에 대한 예측값 $\hat{R}(y_i)$을 추정한다.

4. 반응변수의 중심의 순위에 대한 추정값 $\hat{R}(y_i)$으로부터 유도된 중심 y_i의 예측값 $\hat{y}_i$은 다음과 같다.

$$\hat{y}_i = \begin{cases} y_{(1)}, & \text{if } \hat{R}(y_i) < R(y_{(1)}), \\ y_{(n)}, & \text{if } \hat{R}(y_i) > R(y_{(n)}), \\ y_{(i)}, & \text{if } \hat{R}(y_i) = R(y_{(i)}) \end{cases}$$

이고, 만약 $R(y_{(i)}) < \hat{R}(y_i) < R(y_{(i+1)})$이면

$$\hat{y}_i = y_{(i)} + (y_{(i+1)} - y_{(i)}) \cdot \frac{\hat{R}(y_i) - R(y_{(i)})}{R(y_{(i+1)}) - R(y_{(i)})} \tag{2.5}$$

이다. 여기서 $y_{(i)}$는 집합 $\{y_1, \ldots, y_n\}$에서 크기가 i-번째인 값이다.

반응변수의 두 폭 $l_{y_i}(\alpha)$와 $r_{y_i}(\alpha)$에 대한 추정값은 위에서 제시된 절차를 폭에 대한 관측 자료인

$$\left\{\left(l_{X_{i1}}(\alpha), \ldots, l_{X_{ip}}(\alpha), l_{y_i}(\alpha)\right) : i = 1, \ldots, n\right\} \text{와 } \left\{\left(r_{X_{i1}}(\alpha), \ldots, r_{X_{ip}}(\alpha), r_{y_i}(\alpha)\right) : i = 1, \ldots, n\right\}$$

에 적용함으로 유도할 수 있다. 단, 반응변수의 중심과 다르게 두 폭에 대한 추정값은 양수임을 확인하여야 한다.

2.2. Theil 방법을 이용한 퍼지회귀모형

평면상의 두 점들의 집합에 대한 최상의 적합 직선을 추정하기 위하여 기울기와 절편이 전체 집합의 원소들과 멀리 떨어진 점에 영향을 크게 받지 않는 방법이 Theil에 의하여 연구되었다. 그는 중위수를 이용함으로 오차의 분포에 영향을 받지 않는 방법을 제시하였다. 전체 집합에 속하는 임의의 두 쌍 원소를 선택하여 추정한 절편과 기울기에 대한 중위수를 이용하여 평면상의 직선을 추정하였다. Dietz (1989)는 오차의 구조가 비대칭이거나 혼합된되 이상치가 존재할 경우 Theil 방법이 효율성이 좋음을 보였다. Agee와 Turner (1979) 그리고 Hussain과 Sprent (1983)는 그람-스미트 과정을 이용하여 설명변수를 변환한 후 Theil 방법을 사용하여 중회귀모형 추정하였다.

본 절에서는 퍼지회귀모형을 추정하기 위하여 α-수준집합

$$\left\{\left(X_{i1}(\alpha), \ldots, X_{ip}(\alpha), Y_i(\alpha)\right) : i = 1, \ldots, n\right\}$$

에 Theil 방법을 적용한다. 자료 $\{(x_{i1}, \ldots, x_{ip}, y_i) : i = 1, \ldots, n\}$에 근거하여 중회귀모형

$$y_i = a_0 + a_1 x_{i1} + \cdots + a_p x_{ip}$$

에 포함된 회귀계수 $a_k(k = 0, 1, \ldots, p)$를 추정한 후 반응변수의 중심 y_i을 추정하는 방법은 다음과 같다.

1. 자료 $\{x_1, \ldots, x_p\}$를 그람-스미트 과정을 이용하여 $\{z_1, \ldots, z_p\}$로 변형한다.

$$z_{ik} = \begin{cases} x_{ik}, & \text{if } k = 1, \\ x_{ik} - \sum_{m=1}^{k-1} \dfrac{< x_k, z_m >}{< z_m, z_m >} z_{im}, & \text{if } k > 1, \end{cases}$$

여기서 $x_i = (x_{1i}, \ldots, x_{ni})$이고 $z_i = (z_{1i}, \ldots, z_{ni})$이다.

2. 그람-스미트 과정으로부터 변환된 $\{z_1, \ldots, z_p\}$가 독립변수인 중회귀모형

$$y_i = \theta_0 + \theta_1 z_{i1} + \cdots + \theta_p z_{ip}$$

에 대한 회귀계수 $\theta_k(k = 1, \ldots, p)$를 추정하는 방법은 다음과 같다.

2-1. $\theta_k^{(0)} = 0\,(k = 1, \ldots, p)$이고 $y_i^{(0)} = y_i$라 하자.

2-2. $\delta\theta_k^{(0)}$는 집합 $\{(y_j^{(0)} - y_i^{(0)})/(z_{jk} - z_{ik}) : z_{ik} < z_{jk},\ 1 \le i < j \le n\}$의 중위수이다.

2-3. $\theta_k^{(1)}$과 $y_i^{(1)}$을 각각 $\theta_k^{(0)} + \delta\theta_k^{(0)}$와 $y_i^{(0)} - \delta\theta_k^{(0)} z_{ki}$라 하자.

2-4. 위와 같은 과정을 추정값 $\theta_k^{(j)}$가 수렴할 때까지 반복하여 수렴 값 $\hat{\theta}_k(k = 1, \ldots, p)$을 추정된 회귀계수라 하자. 즉,

$$\delta\theta_k^{(n-1)} = \left\{ \frac{y_j^{(0)} - y_i^{(0)}}{z_{jk} - z_{ik}} - \sum_{s=0}^{n-2} \delta\theta_k^{(s)} : z_{ik} < z_{jk},\ 1 \le i < j \le n \right\},$$

$$y_i^{(n)} = y_i^{(0)} - \sum_{s=0}^{n-1} \delta\theta_k^{(s)} z_{ki}, \quad \theta_k^{(n)} = \sum_{s=0}^{n-1} \theta_k^{(s)}.$$

2-5. $\hat{\theta}_0$는 집합 $\{y_i : i = 1, \ldots, n\}$의 중위수이다.

3. 2단계에서 추정한 $\hat{\theta}_k(k = 1, \ldots, p)$를 사용하여 회귀계수 $a_k(k = 0, 1, \ldots, p)$를 추정하기 위해 그람-스미트 과정을 이용한다.

$$\hat{a}_{p-k} = \begin{cases} \hat{\theta}_p, & \text{if } k = 0, \\ \hat{\theta}_{p-k} - \sum_{m=0}^{k-1} \dfrac{< x_{p-m}, z_{p-k} >}{< z_{p-k}, z_{p-k} >} \hat{\theta}_{p-m}, & \text{if } 1 \le k \le p - 1. \end{cases}$$

그리고 절편에 대한 추정량은 $\hat{a}_0 = \hat{\theta}_0$이다.

특히, 설명변수와 반응변수의 관계가 단순회귀모형

$$y_i = a_0 + a_1 x_{i1}$$

인 경우 기울기의 추정량 $\hat{a}_1$은 $\{(y_j - y_i)/(x_j - x_i) : 1 \le i < j \le n, x_i \ne x_j\}$의 중위수이고 절편의 추정량 $\hat{a}_0$는 $\{y_i - \hat{a}_1 x_i : i = 1, \ldots, n\}$의 중위수이다.

위에서 제시된 절차를 관측된 자료

$$\left\{ \left(l_{X_{i1}}(\alpha), \ldots, l_{X_{ip}}(\alpha), l_{y_i}(\alpha) \right) : i = 1, \ldots, n \right\} \text{와} \ \left\{ \left(r_{X_{i1}}(\alpha), \ldots, r_{X_{ip}}(\alpha), r_{y_i}(\alpha) \right) : i = 1, \ldots, n \right\}$$

에 적용하여 반응변수의 α-수준집합의 왼쪽과 오른쪽의 폭에 대한 추정치 $l_{\hat{y}_i}(\alpha)$와 $r_{\hat{y}_i}(\alpha)$를 유도할 수 있다.

3. 예제

본 절에서는 순위변환방법과 Theil 방법을 이용하여 추정된 퍼지선형회귀모형의 효율성을 비교하는 척도를 소개하고, 본 논문에서 추정된 퍼지회귀모형에 대한 정확성을 다른 방법으로 추정된 모형과 비교한다.

Kim과 Bishu (1998)는 추정된 퍼지회귀모형의 정확성을 비교하기 위해 소속함수에 대한 적분을 사용하였다. 그들은 관측된 퍼지수 Y_i와 추정된 퍼지수 $\hat{Y}_i$의 차를 두 소속함수의 차에 대한 적분으로 다음과 같이 정의하였다.

$$d\left(Y_i, \hat{Y}_i\right) = \int_{-\infty}^{\infty} \left|\mu_{Y_i}(x) - \mu_{\hat{Y}_i}(x)\right| dx.$$

두 퍼지수사이의 차이가 작으면 작을수록 식 (3.1)의 값은 0으로 접근한다. 그러므로 척도 (3.1)의 값이 0에 가까울수록 추정된 퍼지수의 정확성은 높다고 할 수 있다. 또한 관측된 퍼지수의 0-수준집합 $A(0)$의 길이가 길수록 식 (3.1)의 값은 커지는 경향이 있다. 식 (3.1)에 관측된 퍼지수의 넓이를 나누어 주는 것이 추정된 퍼지수의 오차를 잘 반영할 수 있다. 따라서 추정된 퍼지회귀모형에 대한 정확성을 조사할 수 있는 척도를

$$M\left(Y, \hat{Y}\right) = \sum_{i=1}^{n} m\left(Y_i, \hat{Y}_i\right)$$

와 같이 정의할 수 있다. 여기서

$$m\left(Y_i, \hat{Y}_i\right) = \frac{d\left(Y_i, \hat{Y}_i\right)}{\int_{-\infty}^{\infty} \mu_{Y_i}(x)dx}$$

이다. 즉, 관측된 퍼지수와 추정된 퍼지수의 소속함수의 차에 대한 적분의 합인 척도 $M(Y, \hat{Y})$가 0에 가까울수록 추정된 퍼지회귀모형의 정확성은 높다고 할 수 있다. 다음 예제는 추정된 퍼지회귀모형의 정확성을 척도 (3.2)를 사용하여 설명한다.

예제 1. Tanaka 등 (1989)이 제시한 표 1의 자료를 이용하여 여러 사람들이 퍼지회귀모형을 설명하였다. Nasrabadi 등 (2004)은 Tanaka 등(1989), Kim과 Bishu (1998) 그리고 Kao와 Chyu (2003)의 결과를 설명하고 그들의 방법이 다른 방법보다 더 효율적이라는 것을 척도 (3.2)를 이용하여 설명하였다. 표 1에 제시된 자료는 대칭퍼지수이므로 주어진 자료의 중심과 임의의 α-수준 구간의 폭에 대한 회귀방정식을 추청하자. 순위변환방법을 이용하여 추정된 표 1의 자료의 중심과 α-수준 구간의 폭의 순위에 대한 퍼지회귀모형은

$$\hat{R}(y_i) = 2.6 + 0.8\,(R(x_i) - 2.5) \text{ 와 } \hat{R}\left(s_{y_i}(\alpha)\right) = 2.5 + 0.67\,(R(x_i) - 2.5)$$

이다. 위 식과 식 (2.5)을 이용하여 추정된 예측퍼지수와 관측퍼지수에 대한 오차 $m(Y_i, \hat{Y}_i^r)$는 표 1의 오른쪽에 있다. 그리고 Theil 방법을 이용한 결과는

$$\hat{Y}_i^t = (5.1667, 1.85)_T + (1.7917, 0.175)_T x_i$$

이다.

표 1은 Kim과 Bishu (1998), Kao와 Chyu (2003), Nasrabadi 등 (2004)이 제시한 퍼지회귀모형에 대한 오차 $m(Y_i, \hat{Y}_i^{kb})$, $m(Y_i, \hat{Y}_i^{kc})$, $m(Y_i, \hat{Y}_i^{n})$와 Theil 방법을 이용한 퍼지회귀모형에 대한 오차 $m(Y_i, \hat{Y}_i^{t})$를

표 1: Tanaka의 자료와 추정된 모형의 정확성

자료		정확성 척도				
x_i	$Y_i = (y_i, s_{y_i})_T$	$m\left(Y_i, \hat{Y}_i^{kb}\right)$	$m\left(Y_i, \hat{Y}_i^{kc}\right)$	$m\left(Y_i, \hat{Y}_i^{n}\right)$	$m\left(Y_i, \hat{Y}_i^{r}\right)$	$m\left(Y_i, \hat{Y}_i^{t}\right)$
1	$(8.0, 1.8)_T$	2.207	2.789	2.564	1.997	1.798
2	$(6.4, 2.2)_T$	3.025	2.589	2.813	2.857	3.388
3	$(9.5, 2.6)_T$	1.042	0.553	0.718	0.200	1.863
4	$(13.5, 2.6)_T$	2.902	3.363	3.062	1.725	2.074
5	$(13.0, 2.4)_T$	0.850	0.385	0.614	0.713	2.011
	Total	10.026	9.679	9.771	7.492	11.134

표 2: Diamond의 자료

$X_i = \left(x_i, l_{x_i}, r_{x_i}\right)_T$	$Y_i = \left(y_i, l_{y_i}, r_{y_i}\right)_T$
$(21, 4.2, 2.1)_T$	$(4, 0.6, 0.8)_T$
$(15, 2.25, 2, 25)_T$	$(3, 0.3, 0.3)_T$
$(15, 1.5, 2.25)_T$	$(3.5, 0.35, 0.35)_T$
$(9, 1.35, 1.35)_T$	$(2, 0.4, 0.4)_T$
$(12, 1.2, 1.2)_T$	$(3, 0.3, 0.45)_T$
$(18, 3.6, 1.8)_T$	$(3.5, 0.35, 0.7)_T$
$(6, 0.6, 1.2)_T$	$(2.5, 0.25, 0.38)_T$
$(12, 1.8, 2.4)_T$	$(2.5, 0.5, 0.5)_T$

보여준다. 순위변환방법을 이용하여 추정된 퍼지회귀모형의 정확성이 더 좋음을 표 1를 통하여 알 수 있다.

예제 2. Diamond (1988)는 독립변수와 종속변수가 모두 퍼지수인 회귀모형을 설명하기 위하여 표 2의 자료를 제시하였다. Diamond (1988)는 자신에 제시한 자료에 대한 퍼지회귀모형을 두 가지 방법으로 추정하였다. 상수항이 각각 실수인 경우 $\hat{Y}_i^1$와 퍼지수인 경우 $\hat{Y}_i^2$를 제시하였다. 그리고 Choi (2007)는 독립모형을 이용하여 표 2에 대한 퍼지회귀모형을 추정하고 Diamond (1988)의 결과와 비교하였다. 표 2에 제시된 퍼지수는 비대칭이므로 중심과 α-수준 구간의 오른쪽과 왼쪽 폭의 순위에 대한 회귀방정식을 추정하여야 한다. 중심에 대한 결과는

$$\hat{R}(y_i) = 4.5 + 0.885(R(x_i) - 4.5)$$

이고 폭에 대한 결과는 다음과 같다.

$$\hat{R}(l_{y_i}(\alpha)) = 4.5 + 0.836(R(x_i) - 4.5), \qquad \hat{R}(r_{y_i}(\alpha)) = 4.5 + 0.132(R(x_i) - 4.5).$$

표 3에 제시된 순위변환방법에 대한 오차 $m(Y_i, \hat{Y}_i^r)$는 위 식과 식 (2.5)을 이용하여 추정된 예측퍼지수의 결과이다. 중위수를 이용한 Theil 방법의 결과는

$$\hat{Y}_i^t = (0.5 + 0.667x_i,\ 0.0933l_{x_i},\ 0.7857 - 0.0286r_{x_i})_T$$

이다. 표 3은 Diamond (1988)의 두 가지 방법, 순위변환방법과 Theil 방법에 대한 효율성을 보여준다. Theil 방법이 다른 방법보다 정확할 수 있음을 표 3은 보여준다.

두 예제를 통하여 본 논문에서 제시된 순위변환방법과 Theil 방법이 다른 방법보다 더 효율적일 수 있음을 확인하였다. 그러나 본 논문에서 소개된 방법이 최소자승법이나 혹 강인 추정법을 이용하여 추정된 퍼지회귀모형보다 더 효율적인 결과를 보여주는 조건들을 연구하여 한다. 따라서 퍼지회귀모형에 대한 효율성을 비교하는 연구가 계속되어야 한다.

표 3: 추정된 모형의 정확성

y_i	정확성 척도				
	$m\left(Y_i, \hat{Y}_i^1\right)$	$m\left(Y_i, \hat{Y}_i^2\right)$	$m\left(Y_i, \hat{Y}_i^c\right)$	$m\left(Y_i, \hat{Y}_i^r\right)$	$m\left(Y_i, \hat{Y}_i^t\right)$
4	0.484	0.283	0.475	0.757	0.201
3	1.377	1.370	1.059	1.947	0.850
3.5	0.972	1.099	1.042	0.356	1.361
2	1.286	1.545	1.437	1.757	0.775
3	0.890	0.873	0.783	0.284	1.270
3.5	0.359	0.564	0.240	0.448	0.038
2.5	1.413	1.887	1.442	1.472	2.286
2.5	1.094	1.173	0.927	1.492	0.550
Total	7.875	8.794	7.405	8.508	7.339

4. 결론

본 논문에서는 독립변수와 종속변수에 대한 퍼지관계를 표현하는 퍼지회귀모형를 추정하기 위하여 분포무관 추정량으로 알려진 순위변환방법과 Theil 방법을 소개하였다. 관측된 퍼지자료의 α-수준집합을 중심, 오른쪽 폭 그리고 왼쪽 폭으로 표현하였다. 순위변환방법은 관측된 α-수준집합의 각 성분의 순위에 대한 회귀모형을 이용하여 퍼지회귀모형을 추정하였다. 그리고 Theil 방법은 두 변수의 α-수준집합의 각 성분으로 구성된 집합에서 선택한 임의의 두 쌍 중심(혹은 폭)으로부터 계산된 변화율 중 중위수를 중심 (혹은 폭)에 대한 추정량으로 간주하였다. 본 논문에서 제시된 순위변환방법과 Theil 방법이 여러 사람들에 의해서 추정된 퍼지회귀모형보다 더 정확할 수 있음을 예제를 통하여 확인하였다.

앞으로 순위변환방법과 Theil 방법을 이용하여 추정한 퍼지회귀모형에 대한 통계적인 성질을 연구하여 두 방법을 이용한 퍼지회귀모형의 타당성이 더 효율적인 조건을 유도할 필요가 있다.

참고 문헌

Agee, W. S. and Turner, R. H. (1979). Application of robust regression to trajectory data reduction, In *Robustness in Statistics*, Academic Press.

Choi, S. H. (2007). Seperate fuzzy regression with fuzzy input and output, *The Korean Communication in Statistics*, **14**, 183–193.

Diamond, P. (1988). Fuzzy least squares, *Information Sciences*, **46**, 141–157.

Dietz, E. J. (1989) Teaching regression in a nonparametric statistics course, *The American Statistician*, **43**, 35–40.

Hussain, S. S. and Sprent, P. (1983). Non-parametric regression, *Journal of the Royal Statistical Society, Series A*, **146**, 182–191.

Iman, R. L. and Conover, W. J. (1979). The use of the rank transform in regression, *Technometrics*, **21**, 499–509.

Kao, C. and Chyu, C. (2003). Least Squares estimates in fuzzy regression analysis, *European Journal of Operational Research*, **148**, 426–435.

Kao, C. and Lin, P. (2005). Entropy for fuzzy regression analysis, *International Journal of System Science*, **36**, 869-876.

Kim, B. and Bishu, R. R. (1998). Evaluation of fuzzy linear regression models by comparing membership functions, *Fuzzy Sets and Systems*, **100**, 343–352.

Kim, H. K., Yoon, J. H. and Li, Y. (2008). Asymptotic properties of least squares estimation with fuzzy observations, *Information Sciences: An International Journal*, **178**, 439–451.

Kim, K. J. and Chen, H. R. (1997). A comparison of fuzzy and nonparametric linear regression, *Computers & Operations Research*, **24**, 505–519.

Nasrabadi, M. M. and Nasrabadi, E. (2004). A mathematical programming approach to fuzzy linear regression analysis, *Applied Mathematical and Computation*, **155**, 873–881.

Tanaka, H., Hayashi, I. and Watada, J. (1989). Possibilistic linear regression analysis for fuzzy data, *European Journal of Operational Research*, **40**, 389–396.

Tanaka, H., Uejima, S. and Asai, K. (1980). Fuzzy linear regression model, *International Congress Applied Systems and Cybernetics*, **4**, 2933–2938.

Tanaka, H., Uejima, S. and Asai, K. (1982). Linear regression analysis with fuzzy model, *IEEE Transaction on Systems, Man, and Cybernetics*, **12**, 903–907.

Wang, N., Zhang, W. and Mei, C. (2007). Fuzzy nonparametric regression based on local linear smoothing technique, *Information Sciences*, **177**, 3882–3900.

Zadeh, L. A. (1965). Fuzzy sets, *Information and Control*, **8**, 338–353.

2009년 7월 접수; 2009년 8월 채택

Fuzzy Linear Regression Using Distribution Free Method

Jinhee Yoon[a], Seunghoe Choi[1,b]

[a]School of Economics, Yonsei University
[b]School of Liberal arts and Sciences, Korea Aerospace University

Abstract

This paper deals with a rank transformation method and a Theil's method based on an α-level set of a fuzzy number to construct a fuzzy linear regression model. The rank transformation method is a simple procedure where the data are merely replaced with their corresponding ranks, and the Theil's method uses the median of all estimates of the parameter calculated from selected pairs of observations. We also consider two numerical examples to evaluate effectiveness of the fuzzy regression model using the proposed method and of another fuzzy regression model using the least square method.

Keywords: Fuzzy regression model, α-level set, rank transformation method, Theil's method.

[1] Corresponding author: Professor, School of Liberal arts and Sciences, Korea Aerospace University, Koyang, Kyungkido, 412-791, Korea. E-mail: shchoi@kau.ac.kr