Journal of the Korean Data & Information Science Society
2017, **28**(5), 971–980

http://dx.doi.org/10.7465/jkdi.2017.28.5.971
한국데이터정보과학회지

가변계수 측정오차 회귀모형[†]

손인석[1] · 심주용[2]

[1]삼성서울병원 통계자료센터 · [2]인제대학교 통계학과

접수 2017년 8월 29일, 수정 2017년 9월 11일, 게재확정 2017년 9월 13일

요 약

가변계수 회귀모형은 회귀계수의 동적변화를 모형화함으로써 종속변수와 입력변수의 관계에 대한 쉬운 해석이 가능하고 회귀계수의 변동성도 추정할 수 있는 장점을 지니고 있으므로, 여러 과학 분야에서 많은 주목을 받고 있다. 본 논문에서는 입력변수와 출력변수의 오차를 효과적으로 고려한 가변계수 오차모형을 제안한다. 가변계수가 평활변수의 알려지지 않은 형태의 비선형함수이므로 이를 추정하기 위하여 커널 방법을 사용한다. 제안된 모형의 성능에 영향을 미치는 초모수의 최적값을 구하기 위하여 일반화 교차타당성 방법 또한 제안한다. 제안된 방법은 모의자료와 실제자료를 이용한 수치적 연구를 통하여 평가된다.

주요용어: 가변계수 모형, 일반화 교차타당성함수, 측정오차 모형, 커널방법, 평활변수.

1. 서론

Hastie와 Tibshirani (1993)에 의해 소개된 가변계수모형은 회귀계수의 동적인 변화를 모형화할 때 매우 유연하고 강력하다. 가변계수모형은 고전적인 선형회귀모형의 유용한 확장된 형태이며 회귀계수를 단지 상수로 설정하지 않고, 다른 입력변수의 값에 따라 변화하는 함수형태로 가정한다(이때 그 입력변수를 평활변수 혹은 환경변수라고 한다). 평활변수로는 주로 시간, 위치 좌표 등이 사용될 수 있다. 특히 평활변수로 시간이 사용된 경우 시간가변계수모형 (time-varying coefficient model)이라고 한다. 그리고 평활변수의 변화에 영향을 받지 않는 회귀계수가 존재하는 경우 준가변계수모형 (semivarying coefficient model: Zhang등, 2002)이라고 한다. 가변계수모형에서는 선형모형과 같이 회귀계수를 이용하여 종속변수와 입력변수의 관계에 대한 쉬운 해석이 가능하고 회귀계수의 변동성도 추정할 수 있다. 이것이 선형모형과 다른 형태의 비모수모형보다 더 나은 장점이 된다.

예를 들어 임금, 교육연수, 성별변수 (남=1, 여=0)로 이루어진 자료에서 남녀의 평균 임금 차이에 관심이 있다고 가정한다. 주어진 자료의 교육연수에 따른 남녀 임금의 산점도 (Figure 1.1)에서, 교육연수에 따른 평균임금 (실선)을 살펴보면, 교육연수가 낮은 경우 남성과 여성의 평균임금이 차이가 많이 나지만, 교육연수가 증가함에 따라 평균 임금의 차이가 감소함을 알 수 있다

[†] 이 논문은 2015년도, 2017년도 하반기 이공학개인기초연구지원사업의 지원으로 수행된 연구결과임. (NRF-2015R1D1A1A01056582), (NRF-2017R1D1A1B03029792).
[1] (06351) 서울시 강남구 일원동 삼성서울병원, 통계자료센터, 선임 연구원.
[2] 교신저자: (50834) 경남 김해시 어방동, 인제대학교 통계학과, 겸임교수. E-mail: ds1631@hanmail.net

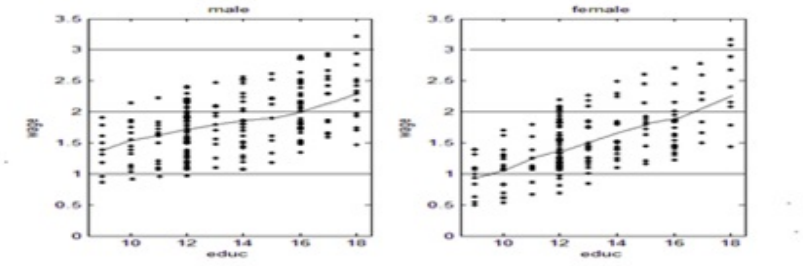

Figure 1.1 Plots of wage versus education

일반적으로 주어진 자료에 대하여 선형회귀모형은 다음과 같이 표현할 수 있다.

$$y \;=\; \beta_0 + \beta_1 x + \beta_2 u + e,$$

여기서 y는 종속변수 (임금), x는 성별변수, u는 교육연수이고, e는 평균이 0이고 유한 분산인 분포를 가지는 오차항이다.

교육연수를 평활변수 (u_i)로 가정하는 가변계수모형은 다음과 같이 표현할 수 있다.

$$y \;=\; \beta_0(u) + \beta_1(u)x + e.$$

자료 $\{u_i \cdot x_i,\ y_i\}_{i=1}^n$를 이용하면, 평활변수 u_i가 주어진 경우 $\boldsymbol{\beta}(u_t) = (\beta_0(u_t), \beta_1(u_t)'$ 의 추정값은 주로 다음과 같은 국소가중다항회귀의 최적화문제의 해로서 구해진다.

$$minL(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} W(u_t - u_i)(y_i - \boldsymbol{\beta}' X_i)^2,$$

여기서 $X_i' \;=\; (1, x_i)$이고 $W(u_t - u_i)$는 $(u_t - u_i)$의 커널함수 (kernel function)이다.

선형회귀모형에서 남녀의 임금 차이는 성별변수에 대응하는 회귀계수 (β_1)의 값으로 나타날 수 있는데 Figure 1.2의 왼쪽 그림과 같이 선형회귀모형에서는 회귀계수 β_1의 추정값은 교육연수의 변화에 따라 상수로 나타난다. 이것은 교육연수가 증가하더라도 남여의 임금 차이가 일정하게(>0) 유지된다는 뜻이다.

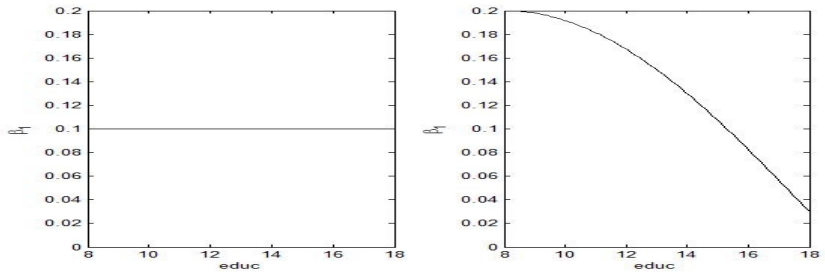

Figure 1.2 Plots of β_1 versus education

가변계수모형에서는 Figure 1.2의 오른쪽 그림과 같이 회귀계수 β_1의 추정값은 교육연수가 증가함에 따라 0으로 감소한다. 이는 교육연수가 증가하면 남여의 임금 차이가 거의 존재하지 않게 된다는 뜻이

다. 따라서 예로서 주어진 자료와 같이 종속변수와 입력변수의 관계가 평활변수의 값에 따라 변화하는 형태의 자료의 분석에서는 선형회귀모형보다 가변계수모형이 많은 장점을 보임을 알 수 있다.

가변계수모형은 최근 통계분야에서 인기를 얻어 비모수 회귀모형, 일반화선형모형 (generalized linear model), 비선형시계열모형, 경시적 (longitudinal) 자료 분석, 생존 (survival) 자료 분석 등에도 응용 범위를 넓혀가고 있다. 기본개념 및 다양한 응용과 연구 분야는 Hoover 등 (1998), Fan과 Zhang (2008)에서 찾을 수 있다. 또한 가변계수를 추정하고 분석하는 방법으로서 국소다항회귀, 커널평활, 다항식스플라인, 평활스플라인 등이 많이 사용되고 있다. 가변계수모형의 추정에 대한 내용은 Fan과 Zhang (2008), Li와 Racine (2010), Lee 등 (2012), Xue와 Qu (2012)에 설명되어 있다.

일반적인 회귀모형에서와는 달리 입력변수의 값을 관찰하는데 오차가 수반된다는 가정에서의 모형을 측정오차모형이라고 하며 이 경우 일반적인 선형회귀분석에서 측정오차가 무시되는 경우 회귀계수의 추정량은 편의추정량이 되고 일치성을 유지하지 못한다 (Fuller, 1987; Caroll 등, 1997). 기본개념 및 다양한 응용과 연구 분야는 Boggs와 Rogers (1990), Van Gorp 등(2000), Hu와 Schennach (2008), Shim (2014)에서 찾을 수 있다.

본 논문에서는 커널기법과 가변계수모형을 측정오차모형에 적용하여 입력변수에 측정오차가 있는 경우 가변계수를 추정할 수 있는 가변계수 측정오차 회귀모형을 제안 한다. 모형 선택의 방법으로는 수정된 형태의 일반화 교차타성함수를 사용한다. 2절에서는 가변계수 측정오차 회귀모형을 제안하고 3절에서는 일반화 교차타당성함수를 사용하는 모형 선택방법을 제안한다. 4절에서는 제안된 방법을 모의자료와 실제자료에 적용하여 다른 방법들과 성능을 비교한다.

2. 가변계수 측정오차 회귀모형

주어진 자료를 $\{\boldsymbol{u}_i, \boldsymbol{x}_i,\ y_i\}_{i=1}^n$라고 표기하기로 한다. 여기서 $\boldsymbol{u}_i \in R_u^d$는 평활벡터, $\boldsymbol{x}_i \in R_x^d$는 입력벡터이고 평균이 $\boldsymbol{x}_i^*$이고 공분산이 $\sigma_e^2 I_{d_x}$ (I_{d_x}는 $d_x \times d_x$ 단위행렬)인 대칭분포에서 관측되었다고 가정한다. 그리고 다음과 같이 종속변수 $y_i \in R$는 $\boldsymbol{u}_i$에 조건적으로 $\boldsymbol{x}_i^*$와 선형적으로 연결되어 있다고 가정한다.

$$y_i = f(\boldsymbol{u}_i,\ \boldsymbol{x}_i^*) + e_i = \sum_{k=0}^{d_x} X_{ik}^* \beta_k(\boldsymbol{u}_i) + e_i, \quad i = 1, \cdots, n, \tag{2.1}$$

여기서 $\boldsymbol{X}_i^* = \begin{pmatrix} 1 \\ \boldsymbol{x}_i^* \end{pmatrix}$, e_i는 평균이 0이고 분산이 σ_e^2인 대칭분포를 독립적으로 따르는 오차항이다. 평활벡터 $\boldsymbol{u}_i$가 주어진 경우 $V(y_i) = \sigma_e^2(1 + \sum_{k=1}^{d_x} \beta_k(\boldsymbol{u}_i)^2)$이다. 정규오차를 가정하고 주어진 자료 $\{\boldsymbol{u}_i, \boldsymbol{x}_i,\ y_i\}_{i=1}^n$를 이용하면 Madansky (1959)가 제안한 방법에 의하면 $\beta_k(\boldsymbol{u}_i)$의 추정값은 다음의 직교잔차제곱합 (sum of squared orthogonal residuals)을 최소화함으로써 구해질 수 있다.

$$\sum_{i=1}^n \left(\frac{y_i - \sum_{k=0}^{d_x} X_{ik} \beta_k(\boldsymbol{u}_i)}{\sqrt{V(y_i)}} \right)^2 = \sum_{i=1}^n \left(\frac{y_i - \sum_{k=0}^{d_x} X_{ik} \beta_k(\boldsymbol{u}_i)}{\sigma_e \sqrt{1 + \sum_{k=1}^{d_x} \beta_k(\boldsymbol{u}_i)^2}} \right)^2. \tag{2.2}$$

식 (2.1)에서 $\beta_k(\boldsymbol{u}_i)$가 평활벡터 $\boldsymbol{u}_i$와 다음과 같이 비선형적으로 연결되어 있다고 가정한다.

$$\beta_k(\boldsymbol{u}_i) = \boldsymbol{\omega}_k' \phi(\boldsymbol{u}_i) + b_k, \quad k = 0, \cdots, d_x,$$

여기서 $\boldsymbol{\omega}_k$는 k번째 입력변수와 $\phi(\boldsymbol{u}_i)$에 대응하는 weight 벡터이고 $\phi(\cdot)$는 비선형 특징사상함수이다.

$(\boldsymbol{\omega}_k,\ b_k)$의 추정을 위하여 다음과 같은 최적화문제를 고려한다.

$$minL = \frac{1}{2}\sum_{k=0}^{d_x}||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n} w_i^{-1}(y_i - \sum_{k=0}^{d_x} X_{ik}\beta_k(\boldsymbol{u}_i))^2 \tag{2.3}$$

$$= \frac{1}{2}\sum_{k=0}^{d_x}||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n} w_i^{-1}(y_i - \sum_{k=0}^{d_x} X_{ik}(\boldsymbol{\omega}_k\phi(\boldsymbol{u}_i) + b_k))^2, \tag{2.4}$$

여기서 $C > 0$ 벌칙상수, $w_i = \left(1 + \sum_{k=1}^{d_x}\beta_k(\boldsymbol{u}_i)^2\right) = \left(1 + \sum_{k=1}^{d_x}(\boldsymbol{\omega}_k\phi(\boldsymbol{u}_i) + b_k)^2\right)$이다.

y_i가 $f(\boldsymbol{u}_i,\ \boldsymbol{x}_i) = \sum_{k=0}^{d_x} X_{ik}\beta_k(\boldsymbol{u}_i)$에 매우 가까운 경우 $f(\boldsymbol{u}_i, \boldsymbol{x}_i) = \sum_{k=0}^{d_x} X_{ik}\beta_k(\boldsymbol{u}_i)(y_i - \sum_{k=0}^{d_x} X_{ik}\beta_k(\boldsymbol{u}_i))^2$는 $\beta_k(\boldsymbol{u}_i)$들의 볼록 (convex)함수가 되므로, y_i가 $f(\boldsymbol{u}_i,\ \boldsymbol{x}_i) = \sum_{k=0}^{d_x} X_{ik}(\boldsymbol{\omega}_k\phi(\boldsymbol{u}_i) + b_k)$에 매우 가까운 경우 최적화문제 (2.3)의 목적함수는 $(\boldsymbol{\omega}_k,\ b_k)$의 볼록함수가 됨을 보일 수 있다. 따라서 최적화문제 (2.3)의 해는 존재하고, $(\boldsymbol{\omega}_k,\ b_k)$의 추정값은 최적화문제 (2.3)의 해 (solution)를 구함으로써 얻어질 수 있다.

따라서 $(\boldsymbol{\omega}_k,\ b_k)$의 추정값을 다음 최적화문제의 해로서 정의한다.

$$min\frac{1}{2}\sum_{k=0}^{d_x}||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n} w_i^{-1}e_i^2. \tag{2.5}$$

제약조건은 $\frac{1}{2}\sum_{k=0}^{d_x}||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n} w_i^{-1}e_i^2,\ \ i = 1, \cdots, n$이다.

위의 최적화문제의 라그랑제 함수는 다음과 같이 구해진다.

$$L = \frac{1}{2}\sum_{k=0}^{d_x}||\boldsymbol{\omega}_k||^2 + \frac{C}{2}\sum_{i=1}^{n} w_i^{-1}e_i^2 - \sum_{i=1}^{n}\alpha_i(e_i - y_i + \sum_{k=1}^{d_x} x_{ik}(\boldsymbol{\omega}_k\phi(\boldsymbol{u}_i) + b_k)),$$

여기서 α_i는 라그랑제 배수이고 최적화 조건 (conditions for optimality)을 이용하면 다음의 결과를 얻을 수 있다.

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\omega}_k} &= \mathbf{0} \ \rightarrow\ \boldsymbol{\omega}_k = \sum_{i=1}^{n} X_{ik}\phi(\boldsymbol{u}_i)\alpha_i, \quad k = 0, \cdots, d_z, \\
\frac{\partial L}{\partial b_k} &= 0 \ \rightarrow\ \sum_{i=1}^{n} X_{ik}\alpha_i = 0, \quad k = 0, \cdots, d_z, \\
\frac{\partial L}{\partial e_i} &= 0 \ \rightarrow\ Cw_i^{-1}e_i - \alpha_i = 0, \quad i = 1, \cdots, n, \\
\frac{\partial L}{\partial \alpha_i} &= 0 \ \rightarrow\ e_i - y_i - \sum_{k=0}^{d_x} X_{ik}(\boldsymbol{\omega}_k\phi(\boldsymbol{u}_i) + b_k) = \text{`}0, \quad i = 1, \cdots, n.
\end{aligned}$$

위의 결과와 Mercer의 조건 (1906)을 이용하면 최적 라그랑제 배수 $(\hat{\alpha}_i)$와 $\hat{b}_k$은 다음의 선형방정식의 해로서 구해진다.

$$\begin{bmatrix} \boldsymbol{X}\boldsymbol{X}' \otimes K(\boldsymbol{u},\ \boldsymbol{u}) + \boldsymbol{W}/C & \boldsymbol{X} \\ \boldsymbol{X}' & \mathbf{0}_{(d_x+1)\times(d_x+1)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0}_{(d_x+1)\times 1} \end{bmatrix}, \tag{2.6}$$

여기서 $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$, $K(\boldsymbol{u},\ \boldsymbol{u}) = \phi(\boldsymbol{u})\phi(\boldsymbol{u})'$, $\boldsymbol{W}$는 w_i로 이루어진 대각행렬, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)'$, $\boldsymbol{b} = (b_0, \cdots, b_{d_x})'$, 그리고 "$\otimes$"는 성분끼리 곱 (component-wise product)을 나타낸다.

따라서 $(\hat{\alpha}_i, \hat{b}_k)$와 커널함수를 이용하면 가변회귀계수의 추정값은 다음과 같이 구해진다.

$$\hat{\beta}_k(\boldsymbol{u}_i) = \sum_{l=1}^{n} X_{lk} K(\boldsymbol{u}_i,\ \boldsymbol{u}_l)\hat{\alpha}_l + \hat{b}_k. \tag{2.7}$$

실제로 w_i가 $\beta_k(u_i)$의 함수이므로 다음과 같은 반복적 방법으로 가변회귀계수의 추정값 $(\hat{\beta}_k(\boldsymbol{u}_i))$을 구해야 한다.

(i) 초기에 $\boldsymbol{W} = I_n$로 놓고, 식 (2.5)와 식 (2.6)을 이용하여 $\hat{\beta}_k(\boldsymbol{u}_i)$를 구한다.
(ii) $w_i = \left(1 + \sum_{k=1}^{d_x} \hat{\beta}_k(\boldsymbol{u}_i)^2\right)$를 이용하여 $\boldsymbol{W}$를 구한다.
(iii) 식 (2.5)와 식 (2.6)을 이용하여 $\hat{\beta}_k(\boldsymbol{u}_i)$를 구한다.
(iv) (ii)와 (iii)을 수렴할 때까지 반복한다.

3. 모형 선택

주어진 모형의 성능 (performance)은 벌칙상수와 커널모수의 값에 영향을 받는다. 모형에 사용되는 커널의 형태가 정해진 경우 모형 선택이란 최적의 벌칙상수와 커널모수의 최적값을 구하는 것이다. 일반적으로 주어진 자료를 훈련용, 시험용 자료로 나누는 k-fold 교차타당성 방법을 사용하는 것이 많이 사용되고, 이때 $k = 1$인 경우 LOO (leave one out)-교차타당성 (cross validation)방법이라고 한다. 일반적으로 모형선택은 교차타당성함수와 격자탐색 (grid search)을 결합하여 이루어진다. 주어진 자료를 $\{\boldsymbol{u}_i,\ \boldsymbol{x}_i,\ y_i\}_{i=1}^n$라고 할 때 식 (2.1)을 이용하여 LOO-교차타당성함수를 다음과 같이 정의한다.

$$CV(\Lambda) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i^{-1}(y_i - \sum_{k=0}^{d_x} X_{ik}\hat{\beta}_k^{(-i)}(\boldsymbol{u}_i))^2,$$

여기서 Λ는 벌칙상수와 커널모수로 이루어진 벡터이고, $\hat{w}_i$는 $w_i = \left(1 + \sum_{k=1}^{d_x} \beta_k(\boldsymbol{u}_i)^2\right)$의 최종 추정값, $\hat{\beta}_k^{(-i)}(\boldsymbol{u}_i)$는 i번째 자료를 이용하지 않고 구한 $\beta_k(\boldsymbol{u}_i)$의 추정값이다.

LOO-교차타당성방법은 주어진 Λ의 값에 대하여 $n(d_x + 1)$개의 $\hat{\beta}_k^{(-i)}(\boldsymbol{u}_i)$가 구해져야하므로 LOO-교차타당성함수를 이용하여 벌칙상수와 커널모수의 최적값을 구하는 것은 계산적으로 매우 비효율적이다. 따라서 우리는 LOO-교차타당성함수를 대체할 교차타당성함수를 고려하여야 한다. leaving-out-one lemma (Craven과 Wahba, 1979)와 Shim과 Hwang (2015)을 이용하면 LOO-교차타당성함수의 근사인 일반화 교차타당성 (generalized cross validation) 함수는 다음과 같이 구해진다.

$$GCV(\lambda) = \frac{n\sum_{i=1}^{n} \hat{w}_i^{-1}(y_i - \sum_{k=0}^{d_x} X_{ik}\hat{\beta}_k(\boldsymbol{u}_i))^2}{(n - trace(H))^2}, \tag{3.1}$$

여기서 $((\boldsymbol{X}\boldsymbol{X}')\otimes K(\boldsymbol{u},\ \boldsymbol{u}))H_0$, $\boldsymbol{X}=(\boldsymbol{X}_1,\ \cdots,\boldsymbol{X}_n)'$, $\boldsymbol{u}=(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n)'$,
$H_0=\begin{pmatrix}\boldsymbol{Z}^{-1}-\boldsymbol{Z}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{Z}^{-1}\boldsymbol{X})^{-1}\boldsymbol{Z}^{-1}\boldsymbol{X}\\ (\boldsymbol{X}'\boldsymbol{Z}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}^{-1}\end{pmatrix}$, $\boldsymbol{Z}=(\boldsymbol{X}\boldsymbol{X}')\otimes K(\boldsymbol{u},\ \boldsymbol{u})+\hat{\boldsymbol{W}}^{-1}/C$, $\hat{\boldsymbol{W}}$는 $\hat{w}_i$로 이루어진 대각행렬이다.

4. 실험 및 결과

첫 번째 예제에서는 모의자료를 이용하여 제안된 가변계수 측정오차모형의 가변계수 추정 성능을 다른 모형 - 가변계수 최소제곱 서포트 벡터 회귀 (varying coefficient least squares vecror regression, VCLSSVR; Shim과 Hwang, 2015), 커널 국소 다항 평활법 (kernel local polynomial smoothing, KLPS; Fan과 Zhang, 2008) - 과 비교한다.

먼저 각 데이터셋의 평활변수 (u_i), 입력변수 (x_i), 가변계수 ($\beta_0(u_i),\ \beta_1(u_i)$), 오차 ($e_i,\ \epsilon_i$) 및 종속변수 ($y_i$)를 다음과 같이 생성한다.

$$u_i\sim U(0,2),\quad x_i^*\sim N(0,4),\quad x_i\sim N(x_i^*,0.25),\quad \beta_0(u_i)=cos(\pi u_i),\quad \beta_1(u_i)=sin(\pi u_i),$$
$$e_i\sim N(0,0.25),\quad y_i=\beta_0(u_i)+\beta_1(u_i)x_i^*+e_i,\quad i=1,\cdots,100,$$
$$\epsilon_i\sim 0.5\times t_4\text{distribution},\quad y_i=\beta_0(u_i)+\beta_1(u_i)x_i^*+\epsilon_i,\quad i=1,\cdots,100.$$

Figure 4.1은 평활변수와 가변계수 ($\beta_0(u_i),\ \beta_1(u_i)$)의 관계를 보여준다. 위와 같은 데이터셋을 100개 생성하여 실험에 이용한다. 가변계수 측정오차모형의 가변계수 추정 성능을 비교하기 위하여 각 데이터셋에서 다음과 같은 평균제곱오차 (MSE)와 평균절대오차 (MAE)을 구한다.

$$MSE_0=\frac{1}{100}\sum_{i=1}^{100}(\beta_0(u_i)-\hat{\beta}_0(u_i))^2,\quad MSE_1=\frac{1}{100}\sum_{i=1}^{100}(\beta_1(u_i)-\hat{\beta}_1(u_i))^2,$$
$$MAE_0=\frac{1}{100}\sum_{i=1}^{100}|\beta_0(u_i)-\hat{\beta}_0(u_i)|,\quad MAE_1=\frac{1}{100}\sum_{i=1}^{100}|\beta_1(u_i)-\hat{\beta}_1(u_i)|.$$

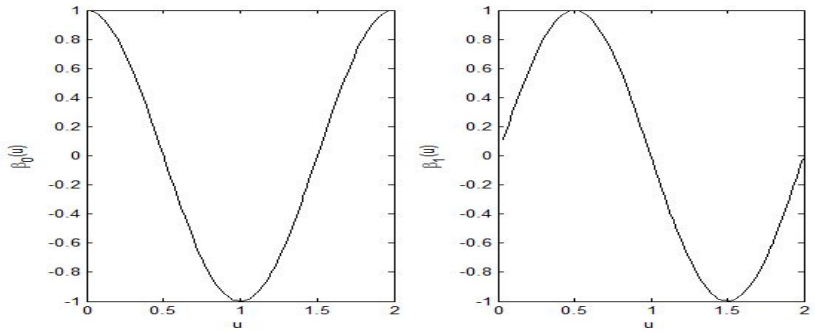

Figure 4.1 Plots of varying coefficients versus smoothing variable in example 1

VCLSSVR과 제안된 모형에서 커널함수로서 가우시안 커널이 사용되었고 벌칙모수 및 커널모수의 최적값은 일반화 교차타당성 함수 (Shim과 Hwang, 2015)과 일반화 교차타당성 함수 (3.1)을 각각 이용

하여 구해진다. Table 4.1은 오차가 정규분포를 따를 때, 가변계수 추정에 대한 100개의 평균제곱오차와 평균절대오차의 평균 및 표준오차를 보여준다. Table 4.2는 오차가 t-분포를 따를 때의 결과를 보여준다. 각 Table에서 진한 표기의 숫자는 세가지 방법에 의한 MSE와 MAE 중 가장 작은 값을 나타내고 괄호안의 숫자는 표준오차를 나타낸다. Table 4.1과 4.2에서 제안한 방법을 적용한 경우 다른 방법에 비해 MSE와 MAE가 작은 값을 가지고, 표준오차도 작으므로 제안된 방법이 더 좋은 추정 결과를 보여준다는 것을 알 수 있다.

Table 4.1 Averages of mean squared errors and mean absolute errors from 100 synthetic datasets with normally distributed responses (Standard errors are in parentheses.)

	$\beta_0(u)$		$\beta_1(u)$	
	MSE	MAE	MSE	MAE
KLPS	0.0366	0.1503	0.0347	0.1568
	(0.0026)	(0.0048)	(0.0011)	(0.0025)
VCLSSVR	0.0233	0.1205	0.0079	0.0701
	(0.0016)	(0.0040)	(0.0004)	(0.0021)
proposed	**0.0229**	**0.1185**	**0.0078**	**0.0694**
	(0.0016)	(0.0042)	(0.0004)	(0.0021)

Table 4.2 Averages of mean squared errors and mean absolute errors from 100 synthetic datasets with responses of distribution (Standard errors are in parentheses.)

	$\beta_0(u)$		$\beta_1(u)$	
	MSE	MAE	MSE	MAE
KLPS	0.0480	0.1713	0.0398	0.1641
	(0.0032)	(0.0060)	(0.0025)	(0.0032)
VCLSSVR	0.0390	0.1535	0.0129	0.0880
	(0.0024)	(0.0049)	(0.0008)	(0.0029)
proposed	**0.0384**	**0.1515**	**0.0123**	**0.0857**
	(0.0026)	(0.0052)	(0.0008)	(0.0028)

두 번째 예제로 Wooldridge (2003)에 있는 526명의 임금자료를 사용한다. 여기서 종속변수로 임금(시간당 달러), 평활변수로 교육연수를, 입력변수로 경력연수를 사용한다. 입력변수로 경력이 측정오차 없이 관측되었지만 실험을 위하여 임금자료의 경력을 x^*로 취급하여 측정오차 있는 입력변수 (x)를 $N(x^*, (0.2 \times \hat{\sigma}(x^*))^2)$에서 생성한다. 제안된 자료의 성능을 확인하기 위하여 자료 (y, u, x^*)를 이용한 VCLSSVR에 의한 가변계수 추정값과 자료 (y, u, x)를 이용한 제안된 방법에 의한 가변계수 추정값을 구해서 비교해 본다. VCLSSVR과 제안된 모형에서 커널함수로서 가우시안 커널이 사용되었고 벌칙모수 및 커널모수의 최적값은 일반화 교차타당성 함수를 이용하여 구해진다. Figure 4.2에서 교육연수가 증가하면 경력연수의 임금에 대한 영향력이 커짐을 알 수 있다. 그리고 제안된 방법이 측정오차가 없는 입력변수를 사용한 경우와 비슷한 정도의 성능을 나타냄을 알 수 있다.

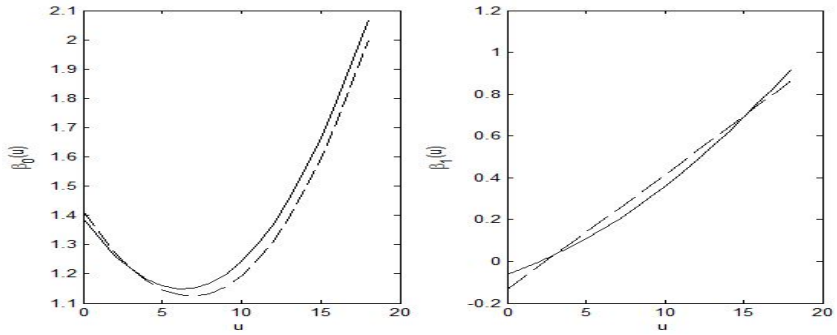

Figure 4.2 Plots of estimates of varying coefficients versus smoothing variable in example 2 (solid=VCLSSVR, dashed=proposed)

5. 결론

본 논문에서는 입력변수에 오차가 포함된 경우 가변계수 모형을 이용하여 가변계수를 추정하는 방법을 다루었다. 우리는 제안된 방법이 주어진 예제에 대한 가변계수를 추정하는데 좋은 결과를 제공한다는 것을 다른 방법들과 비교함으로써 보였다. 제안된 방법은 일반화 교차타당성 함수를 사용하여 leave-one-out 교차 타당성 함수 또는 k 폴드 교차타당성 함수보다 모형 선택이 더 쉽고 빠르다는 것도 알 수 있다. 추후 관련 연구로 입력변수 ($\boldsymbol{x}$)의 평균 ($\boldsymbol{x}^*$)을 가중잔차제곱합 (weighted sum of squared residuals)을 이용하여 추정하는 방법을 고려할 수 있다.

References

Boggs, P. T. and Rogers, J. E. (1990). Orthogonal distance regression. *Contemporary Mathematics*, **112**, 183-194.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1997). *Measurement error in nonlinear models*, Monographs on Statistics and Applied Probability, Chapman & Hall, New York.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numerical Mathematics*, **31**, 377-403.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1**, 179-195.

Fuller, W. A. (1987). *Measurement error models*, Wiley, New York.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: B*, **55**, 757-796.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

Hu, Y. and Schennach, S. M. (2008). Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions using instruments. *Econometrica*, **76**, 195-216.

Lee, Y. K., Mammen, E. and Park, B. U. (2012). Projection-type estimation for varying coefficient regression models. *Bernoulli*, **18**, 177-205.

Li, Q. and Racine, J. S. (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory*, **26**, 1607-1637.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, **54**, 173-205.

Mercer, J. (1909). Function of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society A*, 415-416.

Shim, J. (2014). Quantile regression with errors in variables. *Journal of the Korean Data & Information Science Society*, **25**, 439-446.

Shim, J and Hwang, C. (2015). Varying coefficient modeling via least squares support vector regression. *Neurocomputing*, **161**, 254-259.

Van Gorp, J., Schoukens, J. and Pintelon, R. (2000). Learning neural networks with noisy inputs using the errors-in-variables approach. *IEEE Transactions on Neural Networks*, **11**, 402-414.

Wooldridge, J. M. (2003). *Introductory econometrics: A modern approach*, South-Western Cengage Learning, Mason.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*, **13**, 1973-1998.

Zhang, W., Lee, S. Y. and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis*, **82**, 166-188.

Journal of the Korean Data & Information Science Society
2017, **28**(5), 971–980

http://dx.doi.org/10.7465/jkdi.2017.28.5.971
한국데이터정보과학회지

Varying coefficient model with errors in variables[†]

Insuk Sohn[1] · Jooyong Shim[2]

[1]Statistics and Data Center, Samsung Medical Center
[2]Department of Statistics, Inje University

Received 29 August 2017, revised 11 September 2017, accepted 13 September 2017

Abstract

The varying coefficient regression model has gained lots of attention since it is capable to model dynamic changes of regression coefficients in many regression problems of science. In this paper we propose a varying coefficient regression model that effectively considers the errors on both input and response variables, which utilizes the kernel method in estimating the varying coefficient which is the unknown nonlinear function of smoothing variables. We provide a generalized cross validation method for choosing the hyper-parameters which affect the performance of the proposed model. The proposed method is evaluated through numerical studies.

Keywords: Generalized cross validation function, kernel method, measurement error model, smoothing variable, varying coefficient regression model.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (NRF-2015R1D1A1A01056582),(NRF-2017R1D1A1B03029792).
[1] Senior Researcher, Statistics and Data Center, Samsung Medical Center, Seoul 06351, Korea.
[2] Corresponding author: Adjunct Professor, Institute of Statistical Information, Department of Statistics, Inje University, Kimhae, 50834, Korea. E-mail: ds1631@hanmail.net