Journal of the Korean Data & Information Science Society
2010, **21**(6), 1171–1180

한국데이터정보과학회지

단일지표모형에서 계수 추정방법의 비교†

최영웅[1] · 강기훈[2]

[12]한국외국어대학교 통계학과

접수 2010년 9월 22일, 수정 2010년 11월 10일, 게재확정 2010년 11월 15일

요 약

회귀함수의 비모수적 적합에서 공변량의 차원이 증가함에 따라 추정량의 극한성질이 좋지 않음이 잘 알려져 있다. 이러한 문제점을 극복하기 위한 방법중의 하나는 단일지표모형의 추정을 이용하여 공변량의 차원을 1차원으로 줄이는 것이다. 단일지표모형에서 계수 추정 방법으로는 반복적으로 해를 계산하여 근사치를 구하는 방법인 준모수적 최소제곱법과 비반복적으로 계산하여 구하는 도함수 가중평균법이 있다. 두 추정 방법 모두 모수적인 방법과 같은 수렴비율로 정규근사한다고 알려져 있지만 실질적인 성능에 관한 비교는 이루어지지 않았다. 본 논문에서는 모의실험을 통해 두 방법에 의한 추정치의 분산을 비교하여 어떠한 방법이 좋은지를 파악하고자 한다.

주요용어: 도함수 가중평균법, 비모수적 방법, 준모수적 최소제곱법, 평활량.

1. 서 론

다중선형회귀모형에서는 d-차원의 설명변수와 반응변수 사이의 구조적 관계를 모형화하여 나타내며 각 조건부 기댓값은 선형관계라고 가정한다. 일반화선형회귀모형도 임의의 연결함수를 사용하나 조건부 기댓값이 선형관계라는 가정을 한다. 하지만 각 조건부 기댓값은 항상 선형관계일 수 없으므로 현대의 회귀분석에서는 커널 함수 (kernel function), 스플라인 (spline), 웨이브릿 (wavelet) 등을 이용한 비모수적 방법이 많이 사용되어진다. 그 중에서 커널을 사용한 방법은 다른 비모수적 방법보다 수리적이고 직관적으로 이해가 쉽다는 장점이 있다. 반면, 차원이 증가함에 따라 분산이 급격하게 증가하여 수렴비율의 문제가 생기는데 이를 다차원의 저주라고도 한다 (Silverman, 1986; Härdle, 1990; Hastie와 Tibshirani, 1990). 이러한 문제의 해결을 위해서는 변수선택, 차원의 축소 등등의 방법을 이용하여 다양한 접근법이 연구되고 있다.

Friedman과 Stuetzle (1981)에 의해 제안된 projection pursuit 모형은 설명변수들의 선형 조합인 단변량 함수들의 합으로 회귀 추정을 한다. projection pursuit 모형의 특수한 경우로 설명변수들의 선형조합으로 이루어진 하나의 항을 갖는 단변량 함수로 회귀함수를 추정한 경우의 모형을 계량경제학의 맥락에서는 단일지표모형 (Single index model)으로 불려진다.

Ichimura (1993)는 지표함수 계수의 추정에 있어서 준모수적 최소제곱법을 개발하였다. 이 추정법은 오차의 제곱을 최소로 하는 기준을 기초로 한 연결함수의 비모수적 추정이기 때문에 projection pursuit 모형과 관계가 깊다고 할 수 있다. Hall (1989), Härdle 등 (1993)은 단일지표모형의 추정과 최적의 평

† 이 논문은 한국외국어대학교 2010년 연구비 지원에 의해 이루어졌음.
[1] (449-791) 경기도 용인시 처인구 모현면 왕산리, 한국외국어대학교 대학원 통계학과, 석사과정.
[2] 교신저자: (449-791) 경기도 용인시 처인구 모현면 왕산리, 한국외국어대학교 통계학과, 교수.
E-mail: khkang@hufs.ac.kr

활화방법에 관해 연구하였다. Carroll 등 (1997)은 일반화선형모형, 부분선형모형, 단일지표모형들을 포함하는 일반화된 부분선형 단일지표모형에서의 점근적인 특징을 연구하였다.

단일지표모형에서 계수 추정의 다른 접근 방법은 회귀함수 기울기의 평균인 도함수 가중평균법에 기초한 것이다. 도함수 가중평균법은 Härdle과 Stoker (1989)에 의해 제안되었고, Powell 등 (1989), Newey와 Stoker (1993)는 이 방법의 근사적 성질에 대해 연구하였다. Chu와 Cheng (1995)은 Powell 등 (1989)이 제안한 도함수 가중평균법에서 불완전한 데이터를 대상으로 한 단일지표모형 추정에 대해 연구하였으며 분포함수와 분포의 도함수를 추정할 때 고차 커널을 사용하면 수렴 비율이 더 좋아진다는 사실에 착안하여 고차의 커널들을 사용하였다. 준모수적 최소제곱법에 대한 Ichimura (1993)와 도함수 가중평균법에 대한 Powell 등 (1989)의 이론적인 결과는 Härdle 등 (2004)에 정리되어 있다.

본 논문에서는 단일지표모형을 포함하는 일반화선형 단일지표모형에서 도함수 가중평균법을 사용하여 추정한 결과와 준모수적 최소제곱법으로 추정한 결과를 분산을 기준으로 서로 비교하고자 한다. 지금까지 대표본 성질 외에 실질적인 성능에 대한 비교는 이루어지지 않았다. 본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문에서 비교하고자 하는 두 가지 추정 방법을 소개하였다. 3장에서는 네 가지 모형에 대해 소표본 모의실험을 통해 두 방법의 성능을 비교하였다. 마지막으로 간략한 결론을 4장에서 서술하였다.

2. 단일지표모형에서 계수 추정

단일지표모형은 경제학에서 중요한 역할을 한다. 단일지표는 다른 변수들을 요약하여 하나의 수치로 나타내는 것으로 예를 들어 가격 지표, 성장 지표, 생계비 지표 등이 있다. 이것은 변수 $X_1, \cdots, X_d$들이 갖고 있는 정보를 요약하여 하나의 지표로 나타내어지는 것이다. 단일지표모형의 일반적인 형식은 식 (2.1)과 같다.

$$E(Y|\boldsymbol{X}) = m(\boldsymbol{X}) = g\left\{v_{\boldsymbol{\beta}}(\boldsymbol{X})\right\} \tag{2.1}$$

식 (2.1)에서 $g(\cdot)$는 미지의 연결함수 (link function)이며 $v_{\boldsymbol{\beta}}(\cdot)$는 $\boldsymbol{\beta}$에 의존하는 지표함수로 흔히 선형함수를 사용하고, $\boldsymbol{X}$는 설명변수 벡터로 $\boldsymbol{X} = (X_1, \cdots, X_d)^\top$이다.

단일지표모형의 추정은 두 단계에 걸쳐 이루어진다. 첫 번째 단계에서는 계수 $\boldsymbol{\beta}$를 추정한다. 두 번째 단계에서는 추정된 $\hat{\boldsymbol{\beta}}$을 사용하여 지표값 $\hat{\boldsymbol{v}}_{\boldsymbol{\beta}} = \boldsymbol{X}^\top\hat{\boldsymbol{\beta}}$을 계산하여 얻어진 $\hat{\boldsymbol{v}}_{\boldsymbol{\beta}}$으로 비모수적 회귀함수를 이용하여 미지의 함수 $g(\cdot)$를 추정한다.

단일지표모형에서 계수 $\boldsymbol{\beta}$의 추정은 두 가지 다른 접근 방법이 있다. 첫 번째는 준모수적 최소제곱법 (Semiparametric Least Squares: SLS) 또는 유사최대가능도 추정법 (Pseudo Maximum Likelihood Estimation: PMLE)과 같이 반복적으로 계산하여 근사치를 구하는 방법이다. 준모수적 최소제곱법은 설명변수가 연속형일 때 사용되며 유사최대가능도 추정방법은 이항반응모형의 경우에 사용된다. 두 번째 방법으로는 도함수 가중평균법 (Weighted Average Derivative Estimation: WADE)처럼 바로 추정하는 방법으로 설명변수가 연속형 또는 이산형인 경우 모두 사용할 수 있다. 이들 방법으로 추정된 $\hat{\boldsymbol{\beta}}$은 모두 $\sqrt{n}$의 수렴 비율로 정규근사한다는 사실이 알려져 있다.

2.1. 준모수적 최소제곱법

SLS방법과 SLS에 가중치를 주는 가중 준모수적 최소제곱법 (Weighted Semiparametric Least Square: WSLS)은 Ichimura (1993)에 의해 소개된 방법으로 SLS는 WSLS의 특별한 경우라 할 수 있다. 이 방법에서는 적합된 회귀식으로 설명되지 못하는 자료의 변동을 최소제곱법 형태의 목적함수로

식 (2.2)와 같이 표현한다.

$$Var\{Y|v_{\boldsymbol{\beta}}(\boldsymbol{X})\} = E\left[\{Y - E[Y|v_{\boldsymbol{\beta}}(\boldsymbol{X})]\}^2 \middle| v_{\boldsymbol{\beta}}(\boldsymbol{X})\right] \tag{2.2}$$

식 (2.2)에서 $v_{\boldsymbol{\beta}}(\cdot)$는 $\boldsymbol{\beta}$에 의존하는 지표함수를 나타내며 이 식으로부터 잘 알려진 최소제곱 기준 형태의 다음 목적함수를 생각할 수 있다.

$$\min_{\boldsymbol{\beta}} E\left[Y - E\{Y|v_{\boldsymbol{\beta}}(\boldsymbol{X})\}\right]^2 \tag{2.3}$$

식 (2.3)에서 기댓값을 표본에서 계산되는 값으로 대체하고 이분산의 경우를 대비하여 적절한 가중치 함수를 사용하며, 미지의 조건부 기댓값을 대신하기 위하여 비모수적 추정법을 사용한다. 따라서 WSLS의 추정량은 식 (2.4)와 같이 정의 할 수 있다.

$$\widehat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} \{Y_i - \widehat{m}_{\boldsymbol{\beta}}(\boldsymbol{X_i})\}^2 w(\boldsymbol{X_i}) I(\boldsymbol{X_i} \in \aleph) \tag{2.4}$$

식 (2.4)에서 $\aleph$는 $\boldsymbol{X}$의 support이며 $\hat{m}_{\boldsymbol{\beta}}$는 모수 $\boldsymbol{\beta}$를 안다는 가정하에서 식 (2.1)에 제시된 회귀함수 m의 추정량으로 leave-one-out 방법으로 Nadaraya-Watson 추정량에 가중치를 주어 사용하면 식 (2.5)와 같다.

$$\widehat{m}_{\boldsymbol{\beta}}(\boldsymbol{X_i}) = \frac{\sum_{j\neq i} Y_j K_h\{v_{\boldsymbol{\beta}}(\boldsymbol{X_i}) - v_{\boldsymbol{\beta}}(\boldsymbol{X_j})\} w(\boldsymbol{X_j}) I(\boldsymbol{X_j} \in \aleph_n)}{\sum_{j\neq i} K_h\{v_{\boldsymbol{\beta}}(\boldsymbol{X_i}) - v_{\boldsymbol{\beta}}(\boldsymbol{X_j})\} w(\boldsymbol{X_j}) I(\boldsymbol{X_j} \in \aleph_n)} \tag{2.5}$$

여기서, h는 평활량 (bandwidth)이고 $K_h(\cdot)$는 커널함수로 $K(\cdot/h)/h$를 의미한다. Ichimura (1993)는 식 (2.4)의 추정량에 대하여 적절한 조건을 만족하면 다음과 같은 점근적 성질을 얻을 수 있음을 보였다.

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{l} N(0, \boldsymbol{V^{-1}\Sigma V^{-1}}) \tag{2.6}$$

식 (2.6)에서 $\boldsymbol{V}$와 $\boldsymbol{\Sigma}$는 다음과 같다.

$$\nabla_v = \left.\frac{\partial v_{\boldsymbol{\beta}}(\boldsymbol{X})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}'}, \quad \nabla_m = g'\{v_{\boldsymbol{\beta}}(\boldsymbol{X})\}\left(\nabla_v - \frac{E\{\nabla_v w(\boldsymbol{X})|\, v_{\boldsymbol{\beta}}(\boldsymbol{X}), \boldsymbol{X} \in \aleph\}}{E\{w(\boldsymbol{X})|\, v_{\boldsymbol{\beta}}(\boldsymbol{X}), \boldsymbol{X} \in \aleph\}}\right)$$

$$\boldsymbol{V} = E\left\{w(\boldsymbol{X})\nabla_v\nabla_v^\top \middle| \boldsymbol{X} \in \aleph\right\}, \quad \boldsymbol{\Sigma} = E\left\{w^2(\boldsymbol{X})\sigma^2(\boldsymbol{X})\nabla_v\nabla_v^\top \middle| \boldsymbol{X} \in \aleph\right\}$$

여기서, $\boldsymbol{V}$와 $\boldsymbol{\Sigma}$를 구하기 위해서는 ∇_v의 값을 알아야 하는데 $\boldsymbol{\beta}$를 추정한 관계로 $\boldsymbol{V}$와 $\boldsymbol{\Sigma}$를 추정해야 한다. Ichimura (1993)는 적당한 조건 하에서 아래의 $\hat{\boldsymbol{V}}$과 $\hat{\boldsymbol{\Sigma}}$은 $\boldsymbol{V}$와 $\boldsymbol{\Sigma}$에 대한 일치추정량임을 밝혔다.

$$\widehat{\nabla}_m = \left.\frac{\partial \widehat{m}_{\boldsymbol{\beta}}(\boldsymbol{X_i})}{\partial \boldsymbol{\beta}}\right|_{\widehat{\boldsymbol{\beta}}}$$

$$\widehat{\boldsymbol{V}} = \frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{X_i}) I(\boldsymbol{X_i} \in \aleph_n)\widehat{\nabla}_m\widehat{\nabla}_m^\top$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} w^2(\boldsymbol{X_i}) I(\boldsymbol{X_i} \in \aleph_n)\{Y_i - \widehat{m}_{\boldsymbol{\beta}}(\boldsymbol{X_i})\}^2\,\widehat{\nabla}_m\widehat{\nabla}_m^\top$$

2.2. 도함수 가중평균법

준모수적 가중평균법이 설명변수가 연속형일때만 사용할 수 있다는데에 반하여 도함수 가중평균법은 설명변수가 연속형과 이산형이 섞여 있는 경우에도 사용할 수 있는 방법이다. 설명변수를 $\boldsymbol{X} = (\boldsymbol{T}, \boldsymbol{U})$를 q차원의 연속형 변수 $\boldsymbol{T}$와 p차원의 이산형 혹은 범주형 변수 $\boldsymbol{U}$로 분리하여 생각하자. 문제를 쉽게 하고 비교를 위해 연속형 변수의 모형만을 생각하면

$$E(Y|\boldsymbol{T}) = m(\boldsymbol{T}) = g(\boldsymbol{T}^\top \boldsymbol{\gamma}) \tag{2.7}$$

이다. 여기서 $\boldsymbol{\gamma}$는 $\boldsymbol{T}$의 선형결합으로 지표값을 계산하는데 쓰이는 계수이다. 그리고 도함수의 가중평균 $\boldsymbol{\delta}$를 식 (2.8)과 같이 정의한다.

$$\boldsymbol{\delta} = E\left\{\nabla_m(\boldsymbol{T})w(\boldsymbol{T})\right\} = E\left\{g'(\boldsymbol{T}^\top \boldsymbol{\gamma})w(\boldsymbol{T})\right\}\boldsymbol{\gamma} \tag{2.8}$$

여기서 $\nabla_m(\boldsymbol{T}) = (\partial_1 m(\boldsymbol{T}), \cdots, \partial_q m(\boldsymbol{T}))^\top$은 $m(\cdot)$의 편미분 벡터이고, g'은 $g(\cdot)$의 도함수를 의미하며 $w(\cdot)$은 가중치 함수이다. Powell 등 (1989)은 $\boldsymbol{T}$의 분포 $f(\cdot)$를 가중치 함수로 사용하는 것을 제안하였다. 즉, $w(t) = f(t)$이다. 그래서 이 추정량은 때로는 분포 가중치 ADE (Average Derivative Estimation) 또는 DWADE (Density Weighted Average Derivative Estimation)로 불려진다. 본 논문에서도 가중치 함수를 위와 같이 사용하여 식 (2.8)을 좀 더 자세히 풀어보면 식 (2.9)로 쓸 수 있다.

$$\int_{R^q} \nabla_m(\boldsymbol{t}) f^2(\boldsymbol{t}) dt = \underbrace{\int_R \cdots \int_R}_{q} (\partial_1 m(\boldsymbol{t}), \cdots, \partial_q m(\boldsymbol{t}))^\top f^2(\boldsymbol{t}) dt_1 \cdots dt_q \tag{2.9}$$

식 (2.9)의 식을 부분적분하면 (2.10)으로 나타내어진다.

$$\boldsymbol{\delta} = \begin{pmatrix} -2\int\cdots\int \partial_1 f(\boldsymbol{t}) f(\boldsymbol{t}) m(\boldsymbol{t}) dt_1 \cdots dt_q \\ \\ \vdots \\ \\ -2\int\cdots\int \partial_q f(\boldsymbol{t}) f(\boldsymbol{t}) m(\boldsymbol{t}) dt_1 \cdots dt_q \end{pmatrix} = -2E\left\{\nabla_f(\boldsymbol{t}) m(\boldsymbol{t})\right\} \tag{2.10}$$

만약 $||\boldsymbol{t}|| \rightarrow \infty$이면 $f(\boldsymbol{t})m(\boldsymbol{t}) \rightarrow 0$을 가정할 수 있다. 또한, $m(\boldsymbol{t}) = E(Y|\boldsymbol{t})$와 조건부 기댓값의 성질에 의해 식 (2.11)을 유도할 수 있다.

$$E\left\{\nabla_f(\boldsymbol{T})m(\boldsymbol{T})\right\} = E\left[E\left\{\nabla_f(\boldsymbol{T})Y|\boldsymbol{T}\right\}\right] = E\left\{\nabla_f(\boldsymbol{T})Y\right\} \tag{2.11}$$

그러므로 $\boldsymbol{\delta}$의 추정은 식 (2.12)로 할 수 있다.

$$\widehat{\boldsymbol{\delta}} = -\frac{2}{n}\sum_{i=1}^n \widehat{\nabla}_{f_h}(\boldsymbol{T_i})Y_i \tag{2.12}$$

여기서, $\widehat{\Delta}_{f_h}(\boldsymbol{t})$는 다음의 식으로 나타낼 수 있다.

$$\widehat{\nabla}_{f_h}(\boldsymbol{t}) = \left(-\partial_1 \hat{f}_h(\boldsymbol{t}), \cdots, -\partial_q \hat{f}_h(\boldsymbol{t})\right)^\top,$$

$$\partial_k \hat{f}_h(\boldsymbol{T}) = \frac{1}{nh_1\cdots h_q}\sum_{j=1}^n \partial_k K\left(\frac{t_1 - T_{j1}}{h_1}, \cdots, \frac{t_q - T_{jq}}{h_q}\right)$$

Powell 등 (1989)은 이렇게 추정되어진 $\hat{\boldsymbol{\delta}}$이 식 (2.13)과 같이 $\sqrt{n}$의 속도로 정규분포로 분포수렴함을 밝혔다.

$$\sqrt{n}(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}) \xrightarrow{l} N(0,\boldsymbol{\Sigma}) \tag{2.13}$$

여기서, $\boldsymbol{\Sigma}=4E(\boldsymbol{r}-\boldsymbol{\delta})(\boldsymbol{r}-\boldsymbol{\delta})^\top$, $\boldsymbol{r}=f(\boldsymbol{T})\nabla_m(\boldsymbol{T})-\{Y-m(\boldsymbol{T})\}\nabla_f(\boldsymbol{T})$이다. 준모수적 최소제곱법에서처럼 참값 $\boldsymbol{\Sigma}$대신에 다음의 $\hat{\boldsymbol{\Sigma}}_\delta$를 사용하는 것을 제안하였다 (Powell 등, 1989).

$$\widehat{\boldsymbol{\Sigma}}_\delta = 4\cdot\frac{\sum_i \widehat{\boldsymbol{r}}(z_i)\widehat{\boldsymbol{r}}(z_i)'}{n} - 4\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}', \widehat{\boldsymbol{r}}(z_i)=\frac{-1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\left(\frac{1}{h}\right)^{k+1}K'\left(\frac{\boldsymbol{T}_i-\boldsymbol{T}_j}{h}\right)(y_i-y_j)$$

3. 모의실험

본 논문에서는 네 가지 모형에 대해서 준모수적 최소제곱법과 도함수 가중평균법에 의한 계수추정치의 분산에 대해서 비교하였다. 두 개의 모형에서는 자료의 수가 달라짐에 따라 두 추정 방법의 변화를 살펴보았으며 나머지 두 개의 모형에서는 오차항을 서로 다르게 주었을 경우에 대해서 살펴보았다. 또한 네 가지 모형 모두 연결함수는 항등함수로 하여 모의실험을 하였다. 준모수적 최소제곱법에서는 $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$가 근사적으로 정규분포를 따르는 반면 도함수 가중평균법에서는 $\sqrt{n}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})$가 근사적으로 정규분포를 따르는데 연결함수가 항등함수일 때, $\boldsymbol{\delta}^*=\boldsymbol{\delta}/E[f(x)]$로 하면 $\boldsymbol{\delta}^*=\boldsymbol{\beta}$가 성립하기 때문이다. 각 추정치의 분산을 구하는 것을 200번 반복하여 이 분산의 평균을 비교하였다. 모의실험에서 설명변수는 두 개로 하였다. Härdle과 Stoker (1989)는 설명변수가 d-차원인 경우 $(d+2)$차 커널의 사용을 제안하였다. $(d+2)$차 커널함수 K의 성질은 식 (3.1)과 같으며 적분은 다중적분을 의미한다.

$$\begin{aligned}
&\int K(\boldsymbol{x})d\boldsymbol{x}=1,\\
&\int x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}K(\boldsymbol{x})d\boldsymbol{x}=0, k_1+k_2+\cdots+k_d<d+2,\\
&\int x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}K(\boldsymbol{x})d\boldsymbol{x}\neq 0, k_1+k_2+\cdots+k_d=d+2.
\end{aligned} \tag{3.1}$$

이에 따라 본 논문에서의 커널함수는 정규분포에서 변형되어진 다음 형태의 4차 커널을 사용하였다.

$$K(x)=\left(\frac{3}{2}-\frac{x^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

그리고 비모수적 함수 추정치를 얻기 위해서는 평활량 h를 정해야 하는데, 본 논문에서는 h는 Nadaraya-Watson 추정치를 구할 때 cross-validation 방법으로 추정하여 준모수적 최소제곱법에 사용하고 도함수 가중평균법에는 여기서 구한 평활량 h_s와 h_s주변 몇 개의 값을 사용하여 각 모형에서 모의실험을 시행하였다.

<모형 1> $Y=\beta_1X_1+\beta_2X_2+\epsilon$의 모형에서 설명변수 X_1과 X_2는 $Uniform(0,1)$에서의 난수를 생성하여, 계수 β_1과 β_2는 모두 1로 하였으며 오차항은 $\epsilon_i\sim N(0,0.5^2)$으로 하였다. 표본의 크기를 $n=50,100,200,400$으로 변화시키며 준모수적 최소제곱법 (WSLS)과 평활량의 변화에 따른 도함수 가중평균법 (WADE)에 의한 추정치 $\hat{\beta}_1$과 $\hat{\beta}_2$의 분산을 비교한 결과가 표 3.1과 그림 3.1에 제시되어 있다.

표 3.1을 보면 준모수적 최소제곱법은 n이 커짐에 따라 분산이 많이 감소하고 있음을 볼 수 있으나 도함수 가중평균법의 분산은 n이 커짐에 따라 큰 폭으로 줄지는 않음을 알 수 있다. 하지만, 적절한 평활량 h를 선택하였을 경우 다른 h를 사용할 때 보다 많이 낮아지는 것을 알 수 있다. 그림 3.1은 도함수 가중평균법에서 h가 변화함에 따라 준모수 최소제곱법과 도함수 가중평균법으로 추정된 $\hat{\boldsymbol{\beta}}$에 대한 분산의 비율로 도함수 가중평균법으로 구한 분산을 준모수 최소제곱법으로 구한 분산으로 나누어 준 값이다. 즉, 값이 1보다 크면 준모수적 최소제곱법으로 추정한 $\hat{\boldsymbol{\beta}}$의 분산이 더 작으며, 1보다 작으면 도함수 가중평균법으로 추정한 $\hat{\boldsymbol{\beta}}$의 분산이 더 작음을 의미한다. 실선은 $\hat{\beta}_1$의 분산 비율을 나타내고 점선은 $\hat{\beta}_2$의 분산 비율을 나타낸다. 이 그림에 의하면 $n = 50, 100, 200$인 경우에는 평활량의 넓은 범위에서 도함수 가중평균법으로 구한 분산이 작게 나타남을 알 수 있고, $n = 400$의 경우에는 상대적으로 좁은 범위의 평활량에서 도함수 가중평균법의 분산이 작음을 알 수 있다.

표 3.1 <모형 1>에서 표본의 크기 및 h에 따른 분산

방법	평활량	계수	$n=50$	$n=100$	$n=200$	$n=400$
WSLS	h_s	$\widehat{\beta}_1$	0.47268	0.23511	0.23279	0.05093
		$\widehat{\beta}_2$	0.36600	0.27248	0.20841	0.05008
WADE	$h_s-0.1$	$\widehat{\beta}_1$	0.42768	0.41281	0.40206	0.39254
		$\widehat{\beta}_2$	0.42683	0.41603	0.40699	0.40511
	h_s	$\widehat{\beta}_1$	0.29113	0.21601	0.19443	0.15397
		$\widehat{\beta}_2$	0.29063	0.22047	0.19132	0.14979
	$h_s+0.2$	$\widehat{\beta}_1$	0.27465	0.25290	0.23875	0.22838
		$\widehat{\beta}_2$	0.27470	0.25274	0.24237	0.22912
	$h_s+0.3$	$\widehat{\beta}_1$	0.21102	0.18535	0.12920	0.12939
		$\widehat{\beta}_2$	0.21086	0.18418	0.12727	0.12993
	$h_s+0.5$	$\widehat{\beta}_1$	0.19533	0.16154	0.16031	0.04847
		$\widehat{\beta}_2$	0.18710	0.14586	0.16334	0.04899
	$h_s+0.6$	$\widehat{\beta}_1$	0.14920	0.13984	0.22851	0.10576
		$\widehat{\beta}_2$	0.14801	0.13961	0.20364	0.09487

<모형 2> $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$의 모형에서 설명변수 X_1은 $X_1 \sim N(0,1)$에서, X_2는 $X_2 \sim N(0.25, 2)$에서 난수를 생성하여, 계수 β_1은 0.2, β_2는 0.8로 하였으며 오차항은 <모형 1>과 다르게 등분산이 아닌 $\epsilon_i \sim N\left(0, \sqrt{(X_{1i} - X_{2i})^2}\right)$으로 하였다. 표본의 크기는 앞의 모형과 같이 $n = 50, 100, 200, 400$으로 변화시키며 두 추정법에 의한 $\hat{\beta}_1$과 $\hat{\beta}_2$의 분산을 비교한 결과가 그림 3.2에 제시되어 있다. 표 3.1과 같이 값을 제시할 수 있으나 공간 사정상 생략하였다. 그림 3.1에서와 마찬가지로 실선은 $\hat{\beta}_1$의 분산 비율을 나타내고 점선은 $\hat{\beta}_2$의 분산 비율을 나타낸다.

첫 번째 모형과 비교해 보면 오차항이 등분산이 아니라면 도함수 가중평균법을 사용하는 것이 평활량의 거의 전 범위에서 $\hat{\beta}_1$과 $\hat{\beta}_2$의 분산을 좀 더 줄일 수 있음을 알 수 있다. 다만, 자료의 크기가 커지면 평활량이 작은 짧은 구간에서만 도함수 가중평균법이 우수하고 나머지에서는 준모수적 최소제곱법이 추정치의 분산을 더 작게 한다.

이제 다른 두 개의 모형에서는 표본의 크기가 아닌 오차항을 식 (3.3)과 같이 변화시키면서 모의실험을 시행하였다. 두 모형에서 모두 $n = 200$으로 하였다.

$$
\begin{aligned}
&(i)\ \epsilon_i \sim N(0,1)\\
&(ii)\ \epsilon_i \sim 0.75N(0,1) + 0.25N(0.25,1)\\
&(iii)\ \epsilon_i \sim 0.75N(-0.5,1) + 0.25N(1.5,2.5)
\end{aligned}
\tag{3.2}
$$

<모형 3> $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$에서 설명변수는 $X_1 \sim N(0,1)$, $X_2 \sim Exp(1)$, 계수 β_1과 β_2는 모

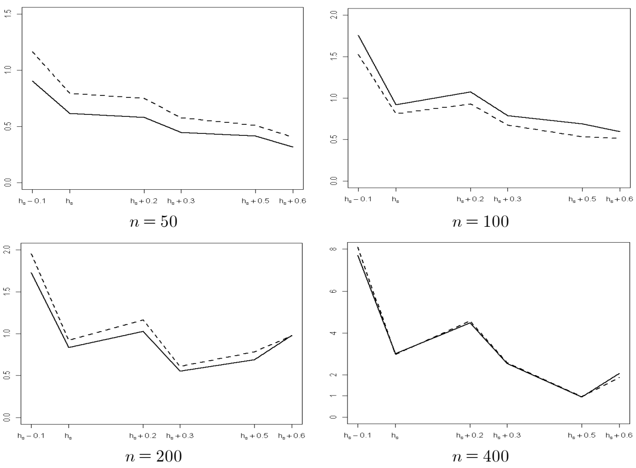

그림 3.1 <모형 1>에서 표본의 크기와 h에 따른 분산 비율

두 1로 하였으며 오차항을 변화시키며 두 방법에 의한 $\hat{\beta}_1$과 $\hat{\beta}_2$의 분산을 비교한 결과는 표 3.2와 그림 3.3의 (a)열에 제시되어 있다. 마찬가지로, 실선은 $\hat{\beta}_1$의 분산비율을 나타내고 점선은 $\hat{\beta}_2$의 분산 비율을 나타낸다. 이 모형에서도 모든 오차항의 경우에 적절한 평활량을 사용하면 도함수 가중평균법이 분산을 더 줄일 수 있음을 알 수 있다.

<**모형 4**> $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$에서 설명변수 X_1과 X_2는 $Exp(1)$, 계수 β_1과 β_2는 모두 1로 하였으며 오차항을 변화시키며 두 방법에 의한 $\hat{\beta}_1$과 $\hat{\beta}_2$의 분산 비율을 비교한 결과는 그림 3.3의 (b)열에 제시되어 있다. 표 3.2와 같이 개별 값들은 공간 사정상 생략하였다. 이 그림에서 보면 오차항이 (iii)인 경우에는 비교적 좁은 구간에서만 도함수 가중평균법의 분산이 더 작음을 알 수 있다.

표 3.2 <모형 3>에서 오차항 및 h에 따른 분산

방법	평활량	계수	(i)	(ii)	(iii)
WSLS	h_s	$\widehat{\beta}_1$	0.87951	0.11261	0.33808
		$\widehat{\beta}_2$	0.72700	0.18290	0.53218
WADE	$h_s - 0.1$	$\widehat{\beta}_1$	1.04332	0.52914	0.63522
		$\widehat{\beta}_2$	1.17187	0.60236	0.44097
	h_s	$\widehat{\beta}_1$	0.74510	0.21654	0.41854
		$\widehat{\beta}_2$	0.73263	0.22862	0.37127
	$h_s + 0.1$	$\widehat{\beta}_1$	0.45656	0.10254	0.18539
		$\widehat{\beta}_2$	0.40552	0.12052	0.17217
	$h_s + 0.2$	$\widehat{\beta}_1$	0.22219	0.15823	0.23900
		$\widehat{\beta}_2$	0.18742	0.18778	0.19196
	$h_s + 0.3$	$\widehat{\beta}_1$	0.32012	0.28700	0.25165
		$\widehat{\beta}_2$	0.29309	0.29963	0.22019

4. 결론 및 고찰

본 논문은 단일지표모형에서 계수 추정방법 중 준모수적 최소제곱법과 도함수 가중평균법을 모의실험을 통하여 추정된 $\hat{\boldsymbol{\beta}}$의 분산을 비교하여 어느 추정방법이 분산의 측면에서 우수한지를 살펴보았다. 모

의실험 결과들을 보면 두 추정방법 중 준모수적 최소제곱법보다는 도함수 가중평균법이 $\hat{\boldsymbol{\beta}}$의 분산이 좀 더 작음을 알 수 있다. 그러나 도함수 가중 평균법은 평활량 h에 따라 추정치의 분산이 다르기 때문에 절대적으로 우수하다고 할 수는 없다. 확인할 수 있는 것은 준모수적 최소제곱법에서 사용되었던 cross-validation으로 구한 평활량 h보다는 좀 더 큰 평활량을 사용해야 도함수 가중평균법에 의한 추정치의 분산이 작아진다는 것이다.

또한, 준모수적 최소제곱법은 설명변수가 연속형일 때만 사용할 수 있으나 도함수 가중평균법의 경우는 설명변수가 연속형에 국한되는 것이 아니라 연속형과 이산형이 함께 있을 때에도 모두 사용할 수 있다는 장점이 있다. 반면, 도함수 가중평균법의 문제점으로는 평활량 h의 선택이 쉽지 않다는 점을 들 수 있겠다. 준모수적 최소제곱방법의 경우 반복적인 수행을 거쳐 결과를 얻어야 하기 때문에 설명변수의 수가 많아지면 도함수 가중평균법에 비해 계산 시간이 훨씬 길어진다는 단점이 있다.

본 논문에서는 두 추정방법 모두 $m(\boldsymbol{X})$의 추정에서 Nadaraya-Watson 추정법을 사용하여 비교하였지만 local polynomial 방법을 사용하여 접근할 수도 있을 것이다. 이 경우엔 계산이 조금 복잡할 것이지만 역시 평활량 h의 선택이 중요한 문제로 남을 것이다. 따라서, 단일지표모형에서 계수를 추정하기 위해 최적 평활량을 구하는 문제도 향후 연구할 가치가 충분하다고 할 것이다.

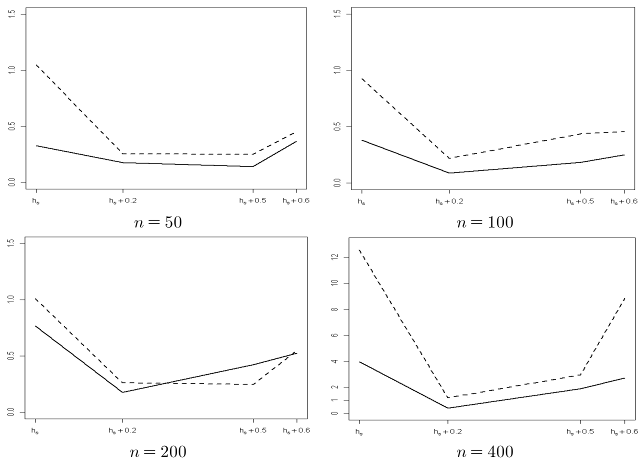

그림 4.1 <모형 2>에서 표본의 크기와 h에 따른 분산 비율

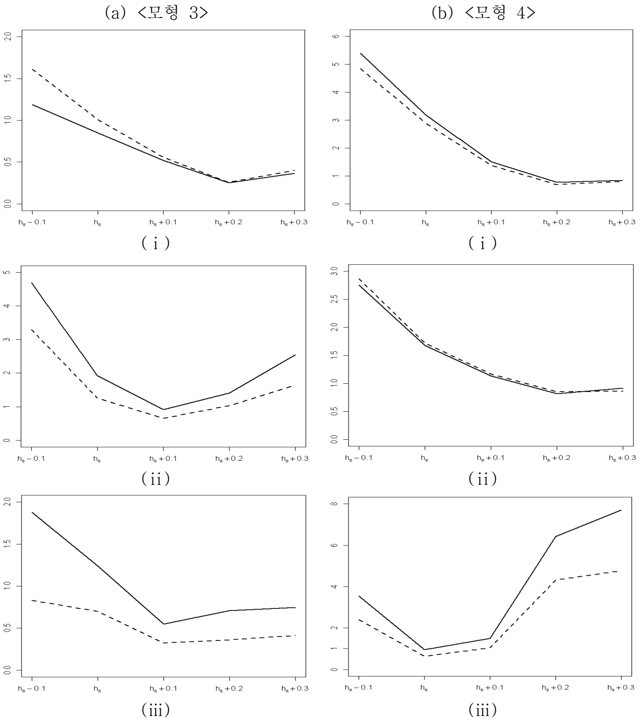

그림 4.2 <모형 3>과 <모형 4>에서 오차항과 h에 따른 분산 비율

참고문헌

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92**, 477-489.

Chu, C. K. and Cheng, K. F. (1995). Nonparametric regression estimates using misclassified binary responses. *Biometrika*, **82**, 315-325

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817-823.

Hall, P. (1989). On projection pursuit regression. *The Annals of Statistics*, **17**, 573-588.

Härdle, W. (1990). *Applied nonparametric regression*, Cambridge University Press.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single index models. *The Annals of Statistics*, **21**, 157-178.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and semiparametric models*, Springer.

Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, **84**, 986-995.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized additive models*, Chapman and Hall.

Ichimura, H. (1993). Semiparametric least square (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71-120.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, **61**, 1199-1223.

Powell, J., Stock, J. and Stoker, T. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403-1430.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Chapman and Hall.

Journal of the Korean Data & Information Science Society
2010, **21**(6), 1171–1180

한국데이터정보과학회지

A comparison on coefficient estimation methods in single index models†

Youngwoong Choi[1] · Kee-Hoon Kang[2]

[12]Department of Statistics, Hankuk University of Foreign Studies

Received 22 September 2010, revised 10 November 2010, accepted 15 November 2010

Abstract

It is well known that the asymptotic convergence rates of nonparametric regression estimator gets worse as the dimension of covariates gets larger. One possible way to overcome this problem is reducing the dimension of covariates by using single index models. Two coefficient estimation methods in single index models are introduced. One is semiparametric least square estimation method, which tries to find approximate solution by using iterative computation. The other one is weighted average derivative estimation method, which is non-iterative method. Both of these methods offer the parametric convergence rate to normal distribution. However, practical comparison of these two methods has not been done yet. In this article, we compare these methods by examining the variances of estimators in various models.

Keywords: Bandwidth, nonparametric method, semiparametric least squares, weighted average derivative estimation.

† This research was supported by the research fund of Hankuk University of Foreign Studies, 2010.
[1] Graduate student, Department of Statistics, Hankuk University of Foreign Studies, Mohyeon, Cheoin-goo, Yongin 449-791, Korea.
[2] Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Mohyeon, Cheoin-goo, Yongin 449-791, Korea. E-mail: khkang@hufs.ac.kr