응용통계연구 제 18권 2호, 2005년, pp. 381-393

분할 역회귀모형에서 차원결정을 위한 점근검정법

박종선[1] 곽재근[2]

요 약

회귀모형에서 필요한 설명변수들의 선형결합들을 탐색하기 위한 방법 중의 하나로 분할역회귀모형을 들 수 있다. 이러한 분할역회귀모형에서 모형에 필요한 설명변수들의 선형결합의 수, 즉 차원을 결정하기 위한 여러 가지의 검정법들이 소개 되었으나 설명변수들의 정규성 가정을 필요로 하거나 다른 제약이 있다. 본 논문에서는 주성분분석에 대한 확률모형을 이용하여 정규성가정을 필요로하지 않으며 분할의 수에 로버스트한 검정법을 소개하고 모의실험과 실제자료에 대한 적용결과를 통하여 기존의 검정법과 비교하였다.

주요용어: 분할역회귀, 차원축소, 잠재변수모형, 점근검정

1. 서론

통계적 예측모형에 속하는 선형 및 일반화 선형모형들은 모형이나 가정에 있어서의 제약 등으로 실제 자료들에 적합시키기에 융통성이 부족한 경우가 많다. 반면, 분할역회귀(SIR: Sliced Inverse Regression, Li; 1991)는 회귀모형에 대한 가정을 최소화하면서 모형에 필요한 설명변수들의 선형결합을 찾는 강력한 기법이라고 할 수 있다. 이 방법은 회귀모형의 형태에 대한 가정이 없이 모형에 하나 이상의 설명변수들의 선형결합들이 필요하다는 가정만을 필요로 하며 효과적으로 모형에 필요한 선형결합들을 탐색할 수 있는 것으로 알려져 있다.

분할역회귀 방법으로 탐색된 선형결합들은 그래픽 방법들을 이용하여 회귀모형에 어떻게 영향을 주는지 판단할 수 있으며 Li(1991)는 분할역회귀와 함께 의미있는 선형결합의 수를 탐색하는 점근 검정법도 소개하였다. 이 검정법은 설명변수들이 정규분포를 따른다는 가정하에 검정통계량이 점근적으로 카이제곱 분포를 따르게 되는 점을 이용하고 있다. Schott(1994)는 설명변수들이 타원형 대칭분포를 따르는 경우 분할 내의 관측치수를 조율모수로 하는 비모수적인 방법을 제안하였다. 1998년에 Velilla는 분할 수와 설명변수에 대한 제약은 없으나 분할 내의 관측치수는 일정한 경우에 대한 검정법을 제시하였으며 같은 해에 Ferre(1998)는 참인 해의 방향과 추정치 사이의 근접 정도를 통하여 차원을 추정하는 모형선택 접근법을 제안하였다. 최근에 Bura와 Cook(2001)은 비 정규성을 갖는 설명변수들의 경우 사용 가능한 가중 카이제곱 검정에 대하여 연구하였다.

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
E-mail:cspark@skku.ac.kr
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 대학원생
E-mail: arilang@naver.com

본 연구에서는 분할 역회귀모형이 일반화 고유값 문제로 표현될 수 있으며 동시에 일반화 고유값 문제는 정규성을 가정한 잠재변수모형으로 나타낼 수 있는 점에 착안하여 설명변수에 대한 분포가정이 없는 경우에 사용 가능한 차원결정 점근 검정법을 제시하고자 한다. Tipping과 Bishop(1999)은 주성분분석의 고유값 분해문제를 위한 잠재변수 모형을 소개하였으며 이를 이용하면 분할역회귀 모형에서의 차원결정문제에 가능도함수를 이용한 점근 검정을 도출할 수 있다. 제시된 검정법은 설명변수에 대한 분포가정이 필요없는 점 외에도 분할의 수에 크게 영향을 받지 않는 것으로 나타났다.

제 2절에서는 분할역회귀모형에 대하여 살펴보고, 분할역회귀에 대한 확률적모형은 제 3절에 포함하였으며 이 모형에 대한 최대가능도추정치를 구하는 EM-알고리즘은 제 4절에서 언급하였다. 모형에 필요한 선형결합의 수인 차원을 결정하기 위한 점근검정법은 제 5절에, 모의 및 실제자료에 대한 비교결과는 제 6절에, 그리고 마지막으로 결론을 포함하였다.

2. 분할 역회귀모형

일변량 반응변수 y와 p차원의 설명변수 $\mathbf{x} = (x_1, x_2, \cdots, x_p)$에 대한 n개의 관측치 $(y_1, \mathbf{x}_1)$, $\cdots, (y_n, \mathbf{x}_n)$를 가지고 있다고 가정하자.

분할역회귀 모형은

$$y = f(\beta^1\mathbf{x}, \beta^2\mathbf{x}, \cdots, \beta^K\mathbf{x}, \epsilon) \tag{2.1}$$

로 표시할 수 있으며 이 때 ϵ는 $\mathbf{x}$와 독립이고 β^i, $i = 1, \cdots, K$는 각각 p차원의 회귀계수들이며 f는 R^{p+1}공간의 임의의 함수이다. 모형에서 알 수 있듯이 분할역회귀 모형은 회귀모형에 특정한 회귀함수를 가정하지 않으며 설명변수들의 선형결합을 하나 이상 모형에 포함시킬 수 있으나 설명변수들이 특정한 조건을 만족하는 경우에만 유용하다. 분할 역회귀 모형에 대한 몇 가지 정의와 제약조건 및 성질들을 살펴보면 다음과 같다.

정의 2.1 모형 (2.1) 하에서 $\beta^1, \cdots, \beta^K$에 의하여 생성되는 β의 공간은 e.d.r.(effective dimension reduction)공간이라 부른다. 그리고 이 공간에서 0 벡터를 제외한 모든 벡터는 e.d.r. 방향이라 부른다.

분할역회귀 모형은 설명변수에 다음과 같은 조건(*Remark 2.1*)이 필요하다. 이 조건은 Li가 "선형조건"이라 하였으며 잘 알려진 정규분포와 타원형 등고선분포들이 이 조건을 만족하는 것으로 알려져 있다.

Remark 2.1: R^p상의 모든 $\mathbf{b}$에 대하여 조건부 기대값 $E(\mathbf{b}\mathbf{x}|\beta^1\mathbf{x}, \cdots, \beta^K\mathbf{x})$는 $\beta^1\mathbf{x}, \cdots, \beta^K\mathbf{x}$의 선형식이 된다. 즉, $c_0, c_1, \cdots, c_K$들을 상수항이라고 할 때 $E(\mathbf{b}\mathbf{x}|\beta^1\mathbf{x}, \cdots, \beta^K\mathbf{x}) = c_0 + c_1\beta^1\mathbf{x} + \cdots + c_K\beta^K\mathbf{x}$가 된다.

정리 2.1 *Remark* 2.1과 모형 (2.1) 하에서 중심화된 회귀선 $E(\mathbf{x}|y) - E(\mathbf{x})$는 $\beta_K\Sigma_{\mathbf{x}}(k = 1, \cdots, K)$ 로부터 파생되는 선형 부분공간(subspace)에 포함된다. 이 때 $\Sigma_{\mathbf{x}}$는 $\mathbf{x}$의 공분산 행렬이다.

설명변수 $\mathbf{x}$를 $\mathbf{z} = \Sigma_{\mathbf{x}}^{-1/2}(\mathbf{x} - E(\mathbf{x}))$로 표준화하면 이론 전개가 간단해지며 역변환을 통하여 다시 원상태로 돌릴 수 있으므로 앞으로의 내용 전개에는 설명변수들을 표준화한 $\mathbf{z}$를 사용하기로 한다.

분할 역회귀는 일차원 변수인 y를 적당한 수의 구간들로 분할하고 각 분할에서 $\bar{y}$와 p차원인 $\mathbf{z}$의 회귀분석을 의미하며 이를 위하여 비모수 방법을 이용하여 역회귀 함수 $E(\mathbf{z}|y)$를 추정하게 된다. 분할 역회귀의 알고리즘을 단계별로 살펴보면 다음과 같다.

단계1: 크기순으로 정렬된 y를 H개의 구간으로 분할한다.

단계2: 각 분할에 대하여 $\mathbf{x}$의 표본평균 $\bar{\mathbf{z}}_h = n_h^{-1} \sum_{(i)\in \text{분할 } h} \mathbf{z}_{(i)}$을 계산한다. 이 때 n_h는 분할 h에서의 관측치의 수이다.

단계3: 분할의 크기에 따른 가중치를 적용하여 분할 내의 $\mathbf{x}$의 평균들에 대한 공분산행렬

$$\hat{\Sigma_\eta} = n^{-1} \sum_{h=1}^{H} n_h \bar{\mathbf{z}}_h \bar{\mathbf{z}}_h'$$

을 계산한다.

단계4: $\hat{\Sigma}_\eta$에 대한 고유값 분할을 이용하여 분할역회귀의 방향

$$\hat{\Sigma}_\eta \hat{\eta}_i = \hat{\lambda}_i \hat{\eta}_i, \quad i = 1, \cdots, k \tag{2.2}$$

을 구한다. 이 때 $\hat{\eta}_i$는 $\eta_i = \beta^i \Sigma_{\mathbf{x}}^{1/2}$에 대한 추정치이며 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_K$ 이다.

여기서 식 (2.2)는 $\hat{\eta}_i$를 성분벡터, 그리고 $\hat{\lambda}_i$를 각각의 성분에 따르는 고유값으로 생각하면 주성분 분석에서의 고유값문제가 됨을 알 수 있다. 주성분분석은 잘 알려진 바와 같이 다변량자료에 대한 차원축소를 목적으로 하는 기법이다. 이 기법은 인자분석(Young, 1940; Whittle, 1952; Anderson, 1963)과 밀접한 관계가 있으며 인자분석은 Lawley(1953)와 Anderson 등(1956)에 의하여 잠재변수모형으로 표현될 수 있음이 알려져 있다. 최근 Tipping과 Bishop(1999)은 기존의 잠재변수 모형을 더 깊게 고찰하여 정규모형의 가정하에서 이 모형의 최대가능도추정량이 주성분분석의 고유값 및 고유벡터들과 같아짐을 보이고 이들 추정치를 구하는 EM 알고리즘을 제시하였다.

3. 확률적 분할 역회귀(Probabilistic SIR: PSIR)

위의 식 (2.1)은 앞에서 언급 하였듯이 p차원의 역회귀 함수 $E(\mathbf{z}|y)$와 이에 연관된 q차원의 잠재변수 $\mathbf{c}$로 이루어진 다음과 같은 잠재변수모형으로 나타낼 수 있다.

$$E(\mathbf{z}|y) = \mathbf{W}\mathbf{c} + \boldsymbol{\epsilon}, \tag{3.1}$$

여기에 분포가정 $\mathbf{c} \sim N(\mathbf{0}, \mathbf{I})$와 오차 ϵ에 대한 추가적인 등방(isotropic) 정규분포, $N(0, \sigma^2\mathbf{I})$를 가정하면 위의 모형 (3.1)과 함께 $\mathbf{c}$가 주어졌을 때의 $E(\mathbf{z}|y)$의 조건부 분포는

$$E(\mathbf{z}|y)|\mathbf{c} \sim N(\mathbf{W}\mathbf{c}, \sigma^2\mathbf{I})$$

이 된다.

잠재변수를 적분하여 얻어지는 z의 주변분포 또한 다음과 같은 정규분포가 된다.

$$E(\mathbf{z}|y) \sim N(\mathbf{0}, \mathbf{\Psi}),$$

이때 $\mathbf{\Psi} = V = Cov(E(\mathbf{z}|y)) = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$ 이다.

z의 주변분포에 대한 공분산행렬 중 $\sigma^2\mathbf{I}$를 일반적인 대각행렬로 바꾸면 위의 모형은 잘 알려진대로 인자분석에 대한 잠재변수모형이 된다. 마지막으로 $\mathbf{W}$와 σ^2에 대한 대수가능도함수는

$$l(\mathbf{W}, \sigma^2) = -\frac{n}{2}\{p\ln(2\pi) + \ln|\mathbf{\Psi}| + \mathrm{tr}(\mathbf{\Psi}^{-1}\mathbf{S})\} \tag{3.2}$$

이 된다. 가능도함수에서

$$\mathbf{S} = \frac{1}{n}\sum_{h=1}^{H} n_h \bar{\mathbf{z}}_h \bar{\mathbf{z}}_h^T = \sum_{h=1}^{H} \hat{\mathbf{z}}_h \hat{\mathbf{z}}_h^T$$

이며 $\hat{\mathbf{z}}_h = \sqrt{n_h/n}\ \bar{\mathbf{z}}$이다.

Tipping과 Bishop은 $\mathbf{W}$와 σ^2에 대한 최대가능도추정치를 구하는 방법으로 EM알고리즘을 제시하였다. 이 방법은 자료에 임의의 결측치가 있는 경우에도 사용할 수 있으며 또한 잠재변수 c에 대한 조건부분포는 베이즈 정리를 이용하면 다음과 같아진다.

$$\mathbf{c}|E(\mathbf{z}|y) \sim N(M^{-1}\mathbf{W}^T E(\mathbf{z}|y), \sigma^2\mathbf{M}^{-1}),$$

분산항에서 $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$이다.

4. EM 알고리즘을 이용한 최대가능도추정법

$\mathbf{W}$와 σ^2에 대한 최대가능도추정치는 EM 알고리즘을 이용하여 구할 수 있다. 이 때 $\mathbf{c}_h$를 EM 알고리즘의 결측부분으로 처리하고 $E(\mathbf{z}|y)$를 일치추정량 $\hat{\mathbf{z}}_h$로 대체한 후 완전한 자료를 $(\hat{\mathbf{z}}_h, \mathbf{c}_h)$으로 놓는다. 관측 및 결측된 자료를 모두 포함하는 전체자료에 대한 대수가능도는 다음과 같다.

$$l_{\hat{\mathbf{z}},\mathbf{c}} = \sum_{h=1}^{H} \ln\{f(\hat{\mathbf{z}}_h, \mathbf{c}_h)\},$$

이때 h 번 째 관측치에 대한 밀도함수는

$$f(\hat{\mathbf{z}}_h, \mathbf{c}_h) = (2\pi\sigma^2)^{-p/2}\exp\left\{-\frac{||\hat{\mathbf{z}}_h - \mathbf{W}\mathbf{c}_h||^2}{2\sigma^2}\right\}(2\pi)^{-q/2}\exp\left\{\frac{||\mathbf{c}_h||^2}{2}\right\}$$

이 된다. EM 알고리즘의 E 및 M-단계를 살펴보면 다음과 같다.

4.1. E-단계

E-단계에서는 전 단계에서 주어진 모수들에 대한 추정치를 이용하여 분포 $f(\mathbf{c}_h|\hat{\mathbf{z}}_h, \mathbf{W}, \sigma^2)$에 대한 $l_{\hat{\mathbf{z}},\mathbf{c}}$의 조건부 기대값을 취한다. 구해진 $l_{\hat{\mathbf{z}},\mathbf{c}}$의 조건부 기대값 $\langle l_{\mathbf{z},\mathbf{c}}\rangle$는 다음과 같다.

$$\begin{aligned}\langle l_{\hat{\mathbf{z}},\mathbf{c}}\rangle &= -\sum_{h=1}^{H}\left\{\frac{p}{2}\ln\sigma^2+\frac{1}{2}\mathrm{tr}\left(\langle \mathbf{c}_h\mathbf{c}_h^T\rangle\right)+\frac{1}{2\sigma^2}\hat{\mathbf{z}}_h^T\hat{\mathbf{z}}_h\right.\\ &\quad\left.-\frac{1}{\sigma^2}\langle \mathbf{c}_h\rangle^T\mathbf{W}^T\hat{\mathbf{z}}_h+\frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{W}^T\mathbf{W}\langle \mathbf{c}_h\mathbf{c}_h^T\rangle\right)\right\}.\end{aligned}$$

식에서 모수들과 관계가 없는 항들은 포함하지 않았으며

$$\langle \mathbf{c}_h\rangle = \mathbf{M}^{-1}\mathbf{W}^T\hat{\mathbf{z}}_h,$$

$$\langle \mathbf{c}_h\mathbf{c}_h^T\rangle = \sigma^2\mathbf{M}^{-1}+\langle \mathbf{c}_h\rangle\langle \mathbf{c}_h\rangle^T,$$

이고 앞에서와 같이 $\mathbf{M}=\mathbf{W}^T\mathbf{W}+\sigma^2\mathbf{I}$ 이다.

4.2. M-단계

M-단계에서는 E-단계에서 구해진 조건부 기대값 $\langle l_{\hat{\mathbf{z}},\mathbf{c}}\rangle$ 을 극대화하는 $\mathbf{W}$와 σ^2를 구한다. 전 단계의 모수들을 이용하여 새롭게 구해진 $\mathbf{W}$의 최대가능도추정량은 다음과 같다.

$$\begin{aligned}\widehat{\mathbf{W}} &= \left\{\sum_{h=1}^{H}\hat{\mathbf{z}}_h\langle \mathbf{c}_h\rangle^T\right\}\left[\sum_{h=1}^{H}\langle \mathbf{c}_h\mathbf{c}_h^T\rangle\right]^{-1}\\ &= \left[\mathbf{SWM}^{-T}\right]\left[\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\mathbf{SWM}^{-T}\right]^{-1}.\end{aligned}$$

추정량의 식에 포함되어 있는 σ^2과 $\mathbf{W}$는 전 단계에서 구해진 추정치를 의미한다. 다음으로 σ^2에 대한 최대가능도추정량은 다음과 같다.

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{Hp}\sum\left\{||\hat{\mathbf{x}}_h||^2-2\langle \mathbf{c}_h\rangle^T\widehat{\mathbf{W}}^T\hat{\mathbf{z}}_h+\mathrm{tr}\left(\langle \mathbf{c}_h\mathbf{c}_h^T\rangle\widehat{\mathbf{W}}^T\widehat{\mathbf{W}}\right)\right\}\\ &= \frac{1}{p}\mathrm{tr}\left(\mathbf{S}-\mathbf{SWM}^{-1}\widehat{\mathbf{W}}^T\right).\end{aligned}$$

위의 추정량에서 $\mathbf{W}$는 이전의 M-단계에서 구해진 추정치이고 $\widehat{\mathbf{W}}$는 현 단계에서 새롭게 구해진 $\mathbf{W}$의 추정치이다. 마지막으로 위의 E 및 M-단계를 수렴할 때까지 반복하여 최대가능도추정치들을 구할 수 있다.

5. 모형의 차원결정

주성분분석에 대한 잠재변수모형에서 차원의 결정은 인자분석에서 모형에 필요한 인자의 수를 결정하는 가설검정과 비슷한 방법을 이용할 수 있다. 우선 모형에 필요한 차원을

q라 놓으면 순차적으로 1부터 $p-1$까지를 귀무가설로 놓고 대립가설은 $q=p$ 또는 공분산의 형태에 아무런 제약이 없도록 하는 가능도비 검정을 고려할 수 있다. 이 때 귀무가설과 대립가설은

$H_0 : \mathbf{\Psi} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$ 단, $q = 1, 2, \cdots, p-1$
$H_1 :$ $\mathbf{\Psi}$는 귀무가설과 다른 형태의 일반적인 양정치 행렬식

이 된다.

앞 절의 식 (3.2)에서 주어진 $\mathbf{W}$와 σ^2에 대한 대수가능도함수와 이들에 대한 최대가능도추정치를 이용하여 가능도비 검정에 대한 가능도비 검정 통계량을 구하면 다음과 같다.

$$-2\ln\Lambda = N\left(\frac{|\widehat{\mathbf{W}}\widehat{\mathbf{W}}^T + \hat{\sigma}^2\mathbf{I}|}{|\mathbf{S}|}\right).$$

이 통계량은 H와 $H-p$가 커짐에 따라 자유도가 $(1/2)\{(p-q)(p-q+1)\}-1$인 χ^2 분포에 점근적으로 수렴하게 된다. 순차적인 가설검정은 모형에 필요한 차원을 q로 두고 $q=1$에서 부터 시작하여 p-값이 정해진 유의수준보다 작으면 q값을 하나씩 증가시켜 최초로 p-값이 유의수준보다 커지는 경우의 q가 모형의 차원이 된다. 참고로 Li의 점근 검정에 대한 자유도는 $(p-q)(H-q-1)$이다.

6. 모의 및 실제 자료에 대한 적용

제안된 점근검정법에 대한 적합성을 알아보기 위하여 모의자료 및 실제자료에 대한 결과를 Li의 검정법과 비교하여 보기로 한다. 모의실험에서 고려한 모형은 세가지 이며 다음의 여러 가지 조합에 대한 결과를 비교하였다. 우선 설명변수의 수는 5개($p=5$)이며 분할의 수는 10, 20, 50, 100(H)이고, 모형에 필요한 선형결합의 수가 1 및 2인 경우를 고려하였다. 표본수는 모두 1000($n=1000$)으로 고정하였으며 모든 경우에 대하여 1000번 반복실험 후 결과를 비교하였다.

모의실험을 위한 자료의 생성과 검정을 위한 알고리즘은 S-plus를 이용하여 구현하였다.

6.1. 모의자료 I: 선형회귀모형

첫번째 모의실험에 사용된 모형은

$$y = 5 + x_1 - 2x_2 + 1.5x_4 + \epsilon$$

이며 여기서 설명변수들과 오차항은 서로 독립인 표준정규분포에서 추출되었다. 총 5개의 설명변수 중 모형에는 3개만 사용되었으며 모형에 필요한 차원, 즉 설명변수들의 선형결합은 하나가 된다. 표 6.1에서 보는 바와 같이 결과에는 검정통계량과 표준편차, p-값들의 평

표 6.1: 모의실험 결과: 선형회귀모형

차원	H	Li 검정				
		χ^2 평균	χ^2 표준편차	p-값 평균	p-값 표준편차	기각율
1	10	32.11	7.99	0.50	0.29	0.055
	20	70.55	12.35	0.53	0.30	0.036
	50	175.13	20.63	0.72	0.27	0.011
	100	330.94	31.34	0.93	0.13	0.000
차원	H	제안된 검정				
		χ^2 평균	χ^2 표준편차	p-값 평균	p-값 표준편차	기각율
1	10	14.24	6.71	0.25	0.26	0.286
	20	10.92	5.33	0.40	0.30	0.137
	50	9.65	4.80	0.47	0.29	0.075
	100	10.14	5.00	0.44	0.29	0.103

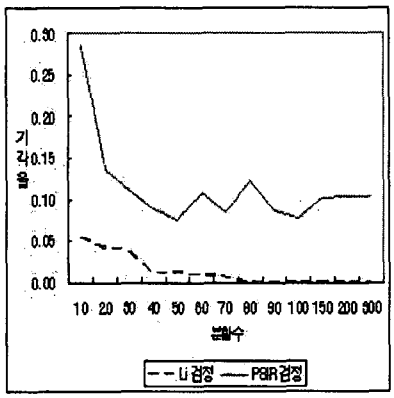

그림 6.1: 분할수에 따른 기각율

균 및 표준편차 그리고 마지막으로 유의수준 5% 하에서 귀무가설에 대한 기각율을 포함하였다.

Li의 점근적 검정은 자유도가 분할수에 의존하므로 검정통계량과 p-값이 분할수에 비례하는 것을 알 수 있다. 그러나 PSIR의 점근적 검정은 분할수에 의존하지 않으므로 분할수 10을 제외하면 거의 비슷한 통계량을 가지며, p-값의 변화도 거의 없는 것을 볼 수 있다.

필요차원을 1로 둔 귀무가설의 경우 전체적으로 Li 검정의 p-값이 높게 나타났으며 기각율의 경우에는 Li의 점근적 검정은 분할수 10에 있어서는 기각율이 실제 유의수준 0.05에 거의 접근하고 있지만, 분할 수가 늘어남에 따라 0에 접근하는 것을 알 수 있다. PSIR의 점

근적 검정은 모든 분할수에서 실제 유의수준 0.05보다 크며 분할수가 커지면 10% 근처로 수렴(그림 6.1)하는 것을 볼 수 있다.

6.2. 모의자료 II: 비선형회귀모형

두번째 모의실험에 사용된 모형은 비선형 회귀모형으로

$$y = (1 + x_1 - 2x_2 + 1.5x_4)^2 + 0.5\epsilon$$

이며 여기서 설명변수들과 오차항은 서로 독립인 표준정규분포에서 추출되었다. 앞의 선형모형과 마찬가지로 총 5개의 설명변수 중 모형에는 3개만 사용되었다. 비록 회귀함수는 선형이 아니지만 이 경우에도 모형의 차원은 1이다.

표 6.2: 모의실험 결과: 비선형회귀모형

		Li 검정				
차원	H	χ^2 평균	χ^2 표준편차	p-값 평균	p-값 표준편차	기각율
1	10	30.36	8.65	0.56	0.31	0.044
	20	60.81	13.60	0.73	0.28	0.013
	50	140.46	24.79	0.95	0.12	0.000
	100	247.23	38.42	1.00	0.01	0.000
		제안된 검정				
차원	H	χ^2 평균	χ^2 표준편차	p-값 평균	p-값 표준편차	기각율
1	10	16.10	8.73	0.22	0.25	0.379
	20	13.06	6.66	0.31	0.29	0.233
	50	13.33	7.10	0.30	0.29	0.243
	100	14.01	6.95	0.27	0.27	0.220

표 6.2의 검정결과를 살펴보면, 선형모형의 경우와 비슷하게 Li의 점근적 검정은 분할수가 증가하면 p-값이 커져 기각율이 0으로 접근하는 경향이 있다. 반면 PSIR의 점근적 검정은 기각율이 선형모형의 경우보다 매우 큰 값들을 나타내고 있으나 분할수가 커질수록 20% 주위에서 수렴(그림 6.2)하는 것을 볼 수 있다.

6.3. 모의자료 III: 유리함수 회귀모형

유리함수 회귀모형에 대한 모의실험에 사용된 모형은

$$y = \frac{x_1 - 2x_2}{0.5 + (1 + x_4 + 2x_5)} + 0.5\epsilon$$

이며 앞의 두 모형과 마찬가지로 설명변수들과 오차항은 서로 독립인 표준정규분포에서 추출되었다. 모형에 필요한 차원은 2가 되며 x_3을 제외한 모든 설명변수들이 사용되었다.

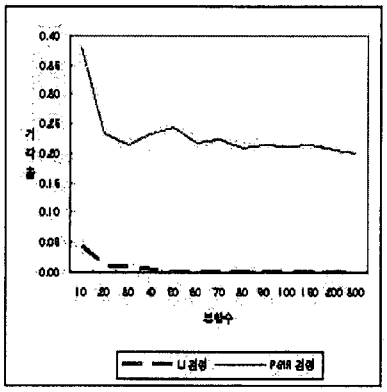

그림 6.2: 분할수에 따른 기각율

표 6.3: 모의실험 결과: 유리함수 회귀모형

차원	H	Li 검정 χ^2 평균	χ^2 표준편차	p-값 평균	p-값 표준편차	기각율
1	10	67.56	14.14	0.01	0.04	0.952
	20	107.50	16.11	0.04	0.08	0.806
	50	183.36	22.40	0.61	0.30	0.038
	100	279.51	30.85	1.00	0.02	0.000
2	10	20.28	6.15	0.53	0.29	0.037
	20	46.12	10.01	0.64	0.29	0.01
	50	108.71	16.20	0.92	0.15	0.001
	100	186.21	22.90	1.00	0.01	0.000
차원	H	제안된 검정 χ^2 평균	χ^2 표준편차	p-값 평균	p-값 표준편차	기각율
1	10	27.04	9.80	0.03	0.08	0.855
	20	24.66	8.85	0.04	0.08	0.813
	50	17.84	8.47	0.16	0.22	0.481
	100	13.94	6.80	0.27	0.27	0.281
2	10	8.55	5.61	0.29	0.28	0.254
	20	6.56	4.31	0.40	0.30	0.141
	50	6.20	4.00	0.41	0.30	0.107
	100	5.89	4.00	0.44	0.30	0.097

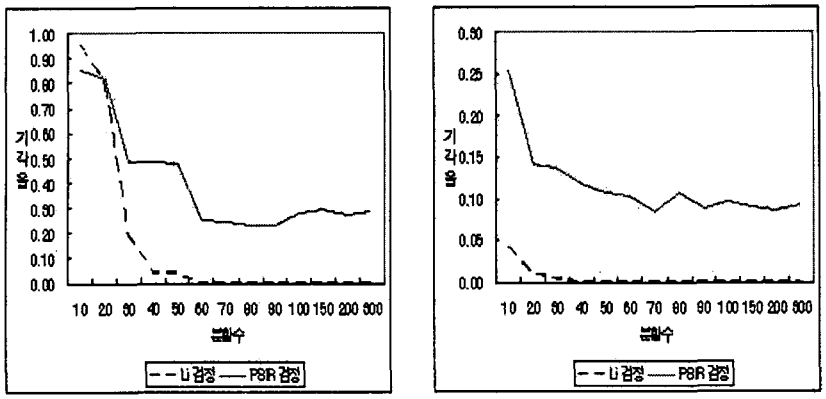

그림 6.3: 왼쪽부터 귀무가설의 차원이 1 및 2인 경우

유리함수 회귀모형에 대한 두 검정방법을 비교한 결과가 표 6.3에 있다. 두 검정 모두 분할수가 작은 경우에 차원을 1로 둔 귀무가설의 기각율이 높은 것으로 나타났으나 Li 검정의 경우 분할수가 큰 경우에는 귀무가설을 적절히 기각하지 못하고 있음을 알 수 있다. 차원을 2로 놓은 검정에서도 Li의 검정은 분할수가 20 이상인 경우 귀무가설을 기각하는 비율이 유의수준 5% 보다 낮았으나 제안된 검정은 앞의 경우들과 마찬가지로 유의수준 5%보다는 높지만 분할수가 커지면 10% 주위의 값으로 수렴(그림 6.2)하는 것을 알 수 있다.

세 가지 모형에 대한 모의실험 결과를 종합하여 보면 제안된 검정의 경우 분할수에 대하여 로버스트하나 접근하는 기각율이 실제 유의수준보다 높게 나타나는 것을 알 수 있다. 기존의 Li 검정은 분할수가 커지면 모든 차원에 대한 검정에서 기각율이 빠른 속도로 0에 접근하는 문제점이 있다.

6.4. 실제자료: 말조개 자료

말조개(Horse Mussel)자료는 뉴질랜드의 Marlborough Sounds섬에서 채집한 82개의 말조개 표본에 대하여 조개살의 무게인 M(Muscle Mass: g)과 4개의 설명변수 W(Shell Width: mm), H(Shell Height: mm), L(Shell Length: mm), S(Shell Mass: g)를 측정한 자료이다.

이 자료에 대해서 분할역회귀의 기본 가정이 만족하는지 알아보기 위해 반응변수 와 설명변수들간의 산점도 행렬을 통해 설명변수 사이의 비선형 정도를 살펴보았다. 그림 6.4의 변수변환 전 자료에 대한 산점도 행렬(왼쪽 그림)에서 비선형 관계를 볼 수 있으며 변수 M과 S에 대한 적절한 변환 후 오른쪽 그림에서 보는 것처럼 설명변수들의 관계가 거의 선형에 가까워 졌음을 알 수 있다. 변환된 자료에 대하여 Li 및 PSIR 검정을 적용한 결과가 표 6.4에 있다.

말조개 자료에 대한 Li의 점근적 검정방법은 분할수가 8인 경우, 검정통계량의 값은 29.66이고 자유도는 18이며, p-값은 0.041로 5% 유의수준에서 귀무가설을 기각할 수 있다.

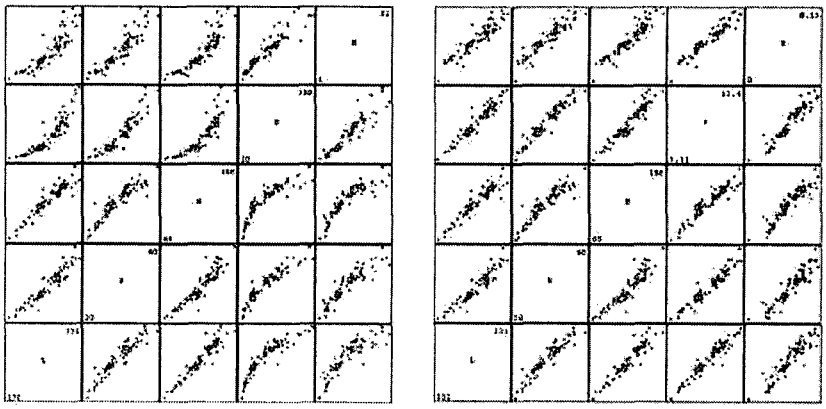

그림 6.4: 왼쪽부터 변수변환($M^{1/3}$, $S^{1/4}$) 전과 후의 설명변수들에 대한 산점도

표 6.4: 말조개 자료 결과

차원	H	Li 검정			제안된 검정		
		χ^2 통계량	자유도	p-값	χ^2 통계량	자유도	p-값
1	8	29.66	18	0.041	9.05	5	0.107
	15	31.75	39	0.789	7.71	5	0.173
	30	48.90	84	0.999	8.19	5	0.146
2	8	9.54	10	0.481	3.07	2	0.215
	15	14.51	24	0.934	3.70	2	0.158
	30	27.15	54	0.999	5.87	2	0.053

그러나 분할수가 15에서는 검정통계량의 값은 31.75고 자유도는 39, p-값은 0.789로 반응변수 를 설명하기 위해서 하나의 선형결합이 필요하다는 결론을 얻게 되며 분할수가 30인 경우의 p-값은 더욱 크게 나타났다. 따라서 말조개 자료에 대한 Li의 점근적 검정은 분할수에 따라서 필요한 선형결합의 수가 다르다는 결론에 이르게 된다. 그리고 분할수에 따라 p-값의 변동이 매우 심하게 나타났다. 그러나 PSIR의 점근적 검정 방법은 분할수의 수에 관계없이 반응변수를 설명하기 위해 필요한 차원이 1라는 귀무가설을 기각할 수 없는 것으로 나타나 앞의 모의실험에서 처럼 분할수에 로버스트한 결과를 얻을 수 있었다.

7. 결론

분할 역회귀 모형에서 필요로 하는 설명변수들의 차원을 결정하기 위한 검정방법으로 Li가 제안한 점근적 방법은 첫째, 설명변수들이 다변량 정규분포를 따른다는 가정을 필요

로 하며, 둘째, 분할수에 따라 p-값의 변동이 큰 것으로 나타났다. 마지막으로 Li의 검정은 분할수가 증가하면 모든 차원에 대한 가설검정에서 p-값이 0으로 빠르게 접근하여 모형에 필요한 차원을 찾는데 실패하는 문제점이 있었다. 다만 설명변수가 정규분포에 가까운 경우에는 분할수가 크지 않도록 조정하여 Li의 검정법을 사용하는 것이 더욱 적합하다고 할 수 있다.

반면 본 논문에서 제안한 확률적 분할 역회귀 모형에 기초한 점근 가능도비 검정법은 무엇보다 설명변수에 대한 분포 가정을 필요로 하지 않으며 연관된 잠재변수 모형에 대한 가능도 함수를 이용할 수 있는 장점이 있다. 더불어 분할의 수에 크게 영향을 받지 않아 로버스트한 것으로 보이나 참의 기각율보다 큰 값으로 수렴하는 경향이 있어 이에 대한 수정이 필요하다고 하겠다.

비록 제안된 검정방법(PSIR 검정)이 Li의 검정방법에 비해 귀무가설에 대한 기각율은 높게 나타나지만, 설명변수의 분포가 알려지지 않은 자료에 적용 가능하며, 분할수와 자료의 수에 영향이 적어 대용량 자료에 효과적으로 적용이 가능한 검정방법이라고 할 수 있다.

참고문헌

Anderson, T. W. (1963). Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics*, **34**, 122-148.

Anderson, T. W., and Rubin, H. (1956). Statistical inference in factor analysis, In J. Newman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume V, U. Cal, Berkeley, 111-150.

Bura, E., and Cook, R. D. (2001). Extending SIR: The weighted chi-square test, *Journal of the American Statistical Association,* **96**, 996-1003.

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods, *Journal of the American Statistical Association,* **93**, 132-140.

Lawley, D. N. (1953). A modified method of estimation in factor analysis and some large sample results, In *Uppsala Symposium on Psychological Factor Analysis*, Number 3 in Nordisk Psykologi Monograph Series, 35-42. Uppsala: Almqvist and Wiksell.

Li, K. C. (1999). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association,* **86**, 316-342.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression, *Journal of the American Statistical Association,* **89**, 141-148.

Tipping, M. E., and Bishop, C. M. (1999). Probabilisitic principal component analysis, *Journal of the Royal Statistical Society, Series B*, **61**, 611-622.

Vellilla, S. (1998). Assessing the number of linear components in a general regression problem, *Journal of the American Statistical Association,* **93**, 1088-1098.

Whittle, P. (1952). On principal components and least square methods of factor analysis, *Skamdinavisk Aktuarietidskrift*, **36**, 223-239.

Young, G. (1940). Maximum likelihood estimation and factor analysis, *Psychometrika*, **6**, 49-53.

[2004년 7월 접수, 2005년 2월 채택]

Asymptotic Test for Dimensionality in Sliced Inverse Regression

Chongsun Park[1] Jae Guen Kwak[2]

ABSTRACT

As a promising technique for dimension reduction in regression analysis, Sliced Inverse Regression (SIR) and an associated chi-square test for dimensionality were introduced by Li (1991). However, Li's test needs assumption of Normality for predictors and found to be heavily dependent on the number of slices. We will provide a unified asymptotic test for determining the dimensionality of the SIR model which is based on the probabilistic principal component analysis and free of normality assumption on predictors. Illustrative results with simulated and real examples will also be provided.

Keywords: Sliced inverse regression, Dimension reduction, Latent variable model, Asymptotic test

1) Professor, Department of Statistics, Sungkyunkwan University, 3-53 Myungryun-dong Jongro-gu, Seoul, 110-745, Korea
E-mail: cspark@skku.edu
2) Graduate Student, Department of Statistics, Sungkyunkwan University, 3-53 Myungryun-dong Jongro-gu, Seoul, 110-745, Korea
E-mail: arilang@naver.com