응용통계연구 제 8권 1호
1995년 3월 pp. 61-74.

다변량 pHd 분석[1)]

이 용 구[2)]

요 약

오늘날에는 컴퓨터를 이용한 다양한 그래프기법의 개발로 자료로부터 정보를 직접적으로 얻는것이 용이하다. 특히 최근에 발표된 R-코드(Cook 과 Weisberg ; 1994)는 다양한 2차원, 3차원 플롯 뿐만아니라 축의 회전과 여러가지 모형에 대한 적합성을 제시하므로 보다 쉽게 자료에 적합한 모형을 시각적으로 분석할 수 있게 하였다. 그러나 그래프는 3차원 이상의 공간을 표현할 수 없기 때문에 하나의 반응변수와 세개이상의 설명변수 사이의 관계를 직접적으로 표현하는 것이 불가능하다. 이와 관련하여 Li(1991,1992)에 의하여 제시된 SIR, pHd방법과 Cook 과 Weisberg(1991)에 의하여 제시된 SAVE는 설명변수들의 선형결합을 이용하여 효과적으로 설명변수들의 차원를 줄이는 방법을 제시하였다. 본 연구에서는 Li에 의하여 제시된 pHd방법을 반응변수가 2개이상인 다변량 반응변수 모형에 적용하는 방법을 연구하였다. pHd방법의 적용에는 많은 계산과정이 요구되는데, 이러한 계산과 다양한 플롯은 R-코드를 이용하였다.

1. 서 론

회귀모형은 $Y, X_{p\times 1}$ 을 각각 반응변수와 설명변수로 할 때, 일반적으로

$$Y = f(X) + \varepsilon \tag{1.1}$$

과 같이 표현할 수 있다. 만약에 위 모형이 k 개의 $p\times 1$벡터들 $\beta_1, \cdots, \beta_k$, $k \le p$ 에 의하여

$$Y = g(\beta_1{}^T X, \cdots \beta_k{}^T X) + \varepsilon \tag{1.2}$$

와 같이 표현될 수 있다면 p개의 설명변수 X는 새로운 k개의 설명변수 $\beta_1{}^T X, \cdots \beta_k{}^T X$ 로 대치될 수 있다. 특히 k가 1 또는 2인 경우에는 2차원 또는 3차원 플롯을 이용하여 Y와 X사이의 관계를 보다 쉽게 분석할 수 있다.

k=p인 경우에는 모형(1.1)과 (1.2)가 동일하므로 $k \le p$에 대하여 위의 조건을 만족하는 $\beta_1, \cdots, \beta_k$ 는 항상 존재한다고 할 수 있다. Li (1991, 1992)는 모형(1.2)를 모형(1.1)의 차원축약모형(dimension reduction model)이라고 정의하였으며, 벡터 $\beta_1, \cdots, \beta_k$ 에 의하여 만들어지는 선형부분공간(linear subspace) $\mathcal{B}$ 를 edr공간(effective dimension reduction space)이라고 정의하였다. 선형부분공간 $\mathcal{B}$의 임의의 기저(basis) B에 대하여 모형(1.2)가 성립하므로 모형(1.2)는 기저 B에 대하여

---

1) 이 논문은 1993년도 중앙대학교 교내 연구비 지원에 의하여 연구 되었음.
2) (156-756) 서울시 동작구 흑석동 221, 중앙대학교 정경대학 응용통계학과.

$$Y = g(B^T X) + \varepsilon \tag{1.3}$$

과 같이 표현할 수 있으며, 부호 " $\perp\!\!\!\perp$ " 가 확률적 독립성을 의미한다고 할 때,

$$Y \perp\!\!\!\perp X \mid B^T X \tag{1.4}$$

와 같이 표현할 수 있다. 모형(1.2)에서 $\beta_1, \cdots, \beta_k$ 의 추정방법으로 Li(1991, 1992)는 역회귀분석방법(Inverse regression method)을 이용한 SIR(Sliced inverse regression)방법과 Hessian 행렬을 이용한 pHd(Principal hessian direction)방법을 제시하였고, Cook 과 Weisberg(1991)는 SIR에 대한 보완적 방법으로 SAVE(Sliced average variance estimate)방법을 제시하였는데, 본 연구에서는 pHd 방법을 다변량반응변수모형에 응용하는 방법을 연구하였다. pHd 방법에서 edr 공간 $\mathcal{B}$의 차원 k의 추정방법과 $\mathcal{B}$의 기저의 추정량 $\hat{B} = (b_1, \cdots b_k)$ 의 추정식은 Li(1992)에 정의되어 있는데, R-코드(Cook 과 weisberg ; 1994)에 이들의 계산 과정이 포함되 있고, 또한 다양한 플롯을 구성할 수 있기 때문에 모의실험및 계산에는 R-코드를 이용하였다.

2. pHd 방법

회귀모형 (1.1)에서 회귀식은 $E(y \mid x) = f(x)$ 와 같다. f(x)가 x에 대하여 2차 미분이 가능하다고 할 때, Hessian행렬 H(x)는

$$H(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right] \tag{2.1}$$

와 같이 정의한다. 다음 보조정리를 통하여 H(x)의 치역공간이 edr공간 $\mathcal{B}$에 속함을 알 수 있다.

보조정리 2.1 회귀모형이 모형(1.2)와 같이 표현될 수 있다면 모든 x값에 대하여 H(x)의 치역공간(range space)은 edr 공간 $\mathcal{B}$에 속한다.

증명 $$H(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right] = \left[\frac{\partial^2 g(B^T X)}{\partial x_i \partial x_j} \right]$$

$$Z_1 = \beta_1^T X, \quad Z_2 = \beta_2^T X, \quad \cdots \quad Z_k = \beta_k^T X,$$

$$g_1 = \frac{\partial g}{\partial z_1}, \quad g_2 = \frac{\partial g}{\partial z_2}, \quad \cdots \quad g_k = \frac{\partial g}{\partial z_k}$$

라고할 때,

$$\frac{\partial g}{\partial z_i} = g_1 \beta_{1i} + g_2 \beta_{2i} + \cdots + g_k \beta_{ki}, \quad i = 1 \cdots p$$

이다. 따라서 일차미분결과는

$$\frac{\partial g}{\partial X} = g_1 \beta_1 + g_2 \beta_2 + \cdots + g_k \beta_k \tag{2.2}$$

와 같으며,

$$g_{ij} = \frac{\partial g_i}{\partial z_j}, \quad i,j = 1, \cdots, k \quad \text{라 할 때,}$$

$$\begin{aligned} H(x) = \; & g_{11}\beta_1\beta_1^T + \cdots + g_{1k}\beta_1\beta_k^T \\ & + g_{21}\beta_2\beta_1^T + \cdots + g_{2k}\beta_2\beta_k^T \\ & + g_{31}\beta_3\beta_1^T + \cdots + g_{3k}\beta_3\beta_k^T \\ & \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ & + g_{k1}\beta_k\beta_1^T + \cdots + g_{kk}\beta_k\beta_k^T \end{aligned} \tag{2.3}$$

이다. 식 (2.3)에 의하여 H(x)의 치역공간은 벡터들 $(\beta_1, \cdots, \beta_k)$ 에 의하여 만들어지는 공간이므로 따라서 H(x)의 치역공간은 $\mathcal{B}$이다.□

보조정리 2.2 모형(1.2)에서 일차미분 결과는 공간 $\mathcal{B}$에 속한다.

증명 식(2.2)에서 $\left[\frac{\partial g}{\partial X}\right]$는 $\beta_1, \cdots, \beta_k$의 선형결합으로 표현되므로 $\left[\frac{\partial g}{\partial X}\right] \in \mathcal{B}$이다.□

Li(1992)는 Hessian행렬의 기대값을 $\overline{H} = EH(x)$라 하고 $\sum_X$ 를 X의 공분산행렬이라고 할 때, pHd공간 $\mathcal{B}$의 기저는 $\overline{H}\sum_X$ 의 0이 아닌 고유근을 갖는 고유벡터들로 구성되며 이와같이 정의된 pHd공간은 선형변환에 불변임을 증명하였다. pHd공간 $\mathcal{B}$의 기저 $B = (\beta_1, \cdots, \beta_k)$의 추정방법에서 Li(1992)는 Stein의 보조정리를 이용하였는데, 이 보조정리는 설명변수 X가 다변량정규분포를 따르는 경우에 이용할 수 있다. 따라서 본 연구에서는 설명변수들이 다변량정규분포를 따른다고 가정하며, 더우기 공간 $\mathcal{B}$가 선형변환에 불변이므로 $X \sim N(0, I_p)$ 인 경우를 대상으로 한다.

3. 다변량 pHd 모형

$Y = (y_1, y_2)^T$, $X = (X_1, \cdots X_p)$, $\varepsilon = (\varepsilon_1, \varepsilon_2)^T$가 각각 반응변수, 설명변수 그리고 오차항이라고 할 때, 다변량회귀모형은

$$Y = f(X) + \varepsilon \tag{3.1}$$

과 같이 표현할 수 있다. pHd방법을 이용하기 위하여 X 와 ε는 각각 다변량 정규분포를 따른다고 가정한다.

k개의 p×1벡터 $(\beta_1, \cdots, \beta_k)$ 에 의하여 모형(3.1)이

$$Y = g(\beta_1^T X, \cdots \beta_k^T X) + \varepsilon \tag{3.2}$$

와 같이 표현될 수 있다면 $B = (\beta_1, \cdots \beta_k)$ 에 대하여

$$Y = g(B^T X) + \varepsilon \tag{3.3}$$

이며,

$$Y \perp\!\!\!\perp X \mid B^T X \tag{3.4}$$

가 된다. 기저가 B인 선형부분공간 $\mathcal{B}$가 Hessian행렬에 의하여 정의될 때 $\mathcal{B}$는 모형(3.1)에 대한 pHd 공간이다. 모형(3.3)은

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_1(B_1^T X) \\ g_2(B_2^T X) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \tag{3.5}$$

와 같이 표현할 수 있다. 따라서 y의 각각에 대하여

$$y_1 \perp\!\!\!\perp X \mid B_1^T X \tag{3.6}$$

$$y_2 \perp\!\!\!\perp X \mid B_2^T X \tag{3.7}$$

인 $B_1 = (\beta_{11}, \cdots \beta_{1k_1})$, $k_1 \le p$ 와 $B_2 = (\beta_{21}, \cdots \beta_{2k_2})$, $k_2 \le p$ 가 존재하며, $\mathcal{B}_1$과 $\mathcal{B}_2$를 각각 B_1과 B_2를 기저로 갖는 선형부분공간이라고 할 때, 특정조건하에서 $\mathcal{B}_1$과 $\mathcal{B}_2$를 이용하여 $\mathcal{B}$를 구할 수 있다.

보조정리 3.1 $X \perp\!\!\!\perp \varepsilon$ 이고 $y_1 \perp\!\!\!\perp y_2 \mid B^T X$ 이면 $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ 이다.

증명 식(3.4)에 의하여 $Y \perp\!\!\!\perp X \mid B^T X$ 이고, f 를 y의 확률분포함수라고할 때 조건에 의하여

$$f(y_1, y_2 \mid B^T X) = f_1(y_1 \mid B^T X) \cdot f_2(y_2 \mid B^T X) \tag{3.8}$$

이다. 또한 $y_1 \perp\!\!\!\perp X \mid B_1^T X$ 이고 $y_2 \perp\!\!\!\perp X \mid B_2^T X$ 이므로 식 (3.8)은

$$\begin{aligned} & f_1(y_1 \mid B_1^T X) \cdot f_2(y_2 \mid B_2^T X) \\ &= f(y_1, y_2 \mid B_1^T X, B_2^T X) \\ &= f(y_1, y_2 \mid (B_1 \cup B_2)^T X) \end{aligned}$$

가 되어 $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ 이다. □

그러나 $X \perp\!\!\!\perp \varepsilon$가 아닌 경우에는 다음예에서와 같이 $\mathcal{B}_1 \cup \mathcal{B}_2 \neq \mathcal{B}$이다.

예 3.1 모형이

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r(X) \\ r(x) & 1 \end{bmatrix} \right] \tag{3.9}$$

이면 $\mathcal{B}_1 = \{0\}, \mathcal{B}_2 = \{0\}$ 이나 $\mathcal{B} = \{r(X)\}$ 이다.

세 부분공간 $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}$ 사이의 관계를 다음예를 통하여 설명해 보기로 한다.

예 3.2 $Y = (y_1, y_2)^T, X = (X_1, X_2, X_3, X_4)$ 이고 $X \perp\!\!\!\perp \varepsilon$ 인 모형

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_1(B_1^T X) \\ g_2(B_2^T X) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

에서

$$B_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad , \quad B_2 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

와 같을때

$$y_1 \perp\!\!\!\perp X \mid B_1^T X \Rightarrow y_1 \perp\!\!\!\perp X \mid (X_1 + X_2, X_3)$$

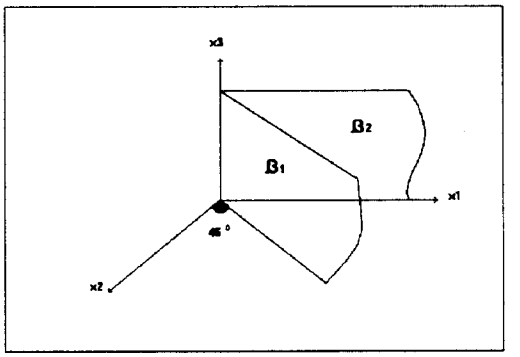
$$y_2 \perp\!\!\!\perp X \mid B_2^T X \Rightarrow y_2 \perp\!\!\!\perp X \mid (X_2 + X_3, X_2)$$

이다. 따라서 $(y_1, y_2) \perp\!\!\!\perp X \mid (X_1 + X_2, X_2, X_3)$ 이며,

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

이고, $\beta = \beta_1 \cup \beta_2$ 임을 알 수 있다.

이 관계를 X_4를 제외한 3차원 공간상의 그림으로 표현하면 [그림 3.1]



[그림 3.1] β_1과 β_2공간

과 같으며, β는 β_1과 β_2의 선형결합으로 이루어지는 공간이다.

여기에서 문제점은 두 선형부분공간 β_1과 β_2 의 기저 B_1, B_2 의 추정량 $\hat{B}_1, \hat{B}_2$ 를 이용하여 어떻게 공간 β의 기저 B를 추정할 수 있는가의 문제이다.

4. 다변량 pHd공간의 추정

식 (3.6)과 (3.7)에서 H_1과 H_2 를 각각

$$H_1 = \left[\frac{\partial^2 g_1(X)}{\partial X_i \partial X_j} \right], \quad H_2 = \left[\frac{\partial^2 g_2(X)}{\partial X_i \partial X_j} \right]$$

와 같이 정의하면 H_1과 H_2는 각각 y_1과 y_2에 대한 Hessian행렬이다. $\overline{H}_1$과 $\overline{H}_2$를 각각 H_1과 H_2의 기대값이라고 할때, β_1과 β_2는 각각 $\overline{H}_1$과 $\overline{H}_2$의 0이 아닌 고유근을 갖는 고유벡터들에 의하여 만들어지는 선형부분공간이라고 할 수 있다. $X \perp\!\!\!\perp \varepsilon$인 경우 세 부분공간 β_1, β_2, β 사이의 관계는 다음과 같이 생각할 수 있다.

(i) $\beta_1 = \beta_2$ 이면 $\beta = \beta_1 = \beta_2$,

(ii) $\beta_1 \subset \beta_2$ 이면 $\beta = \beta_2$,

(iii) $\beta_1 \supset \beta_2$ 이면 $\beta = \beta_1$

(iv) $\beta_1 \perp \beta_2$ 이면 $\beta = \beta_1 \oplus \beta_2$

(v) $\beta_1 \not\subset \beta_2$ 이고 $\beta_1 \not\supset \beta_2$ 이면 $\beta = \beta_1 \cup \beta_2$

즉, β, β_1, β_2 사이의 관계는 위와 같이 5가지 경우로 나누어 생각할 수 있는데, 단지 각 선형공간의 기저의 추정량을 이용하여 두 선형공간 β_1 과 β_2 사이의 관계에 대한 검정을 실시하는것은 불가능한 것으로 판명되었다. 이의 해결방법으로 본 논문에서는 두 선형공간의 동일성, $\beta_1 = \beta_2$ 에 대하여는 반응변수 Y_1 과 Y_2 사이의 관계를 이용하여 증명하고자 하였고, 두 부분공간이 서로 직교하는가와 한 공간이 다른공간의 부분집합 인가에 대하여는 정사영(Orthogonal projection)과 정준상관분석방법(Canonical correlation methods)을 이용하였다.

정리 4.1 $K_1 = K_2 = 1$ 일때, $\{ \beta_1 = \beta_2 \}$의 필요충분조건은 $\{ y_2 = h(y_1) + \delta \}$ 이다. 여기에서 δ 는 오차항으로 $X \perp\!\!\!\perp \delta$이며, h는 2차미분이 가능한 함수이다.

증명 $" \Rightarrow "$ $\beta_1 = \beta_2 = \beta$ 이므로

$$y_1 = g_1(B^T X) + \varepsilon_1$$
$$y_2 = g_2(B^T X) + \varepsilon_2$$

이다. 따라서 $B^T X = g^{-1}(y_1 - \varepsilon_1)$ 이 되며,

$$y_2 = g_2(g_1^{-1}(y_1 - \varepsilon_1)) + \varepsilon_2$$
$$= h(y_1) + \delta$$

이다.

" ← " 벡터 $\beta_1,\cdots,\beta_k$ 로 만들어지는 선형부분공간을 $\mathcal{L}(\beta_1,\cdots,\beta_k)$ 로 표현하기로 한다. 선형부분공간 $\mathcal{B}_1$의 기저를 $B_1=[\beta_1]$이라고 할때, [보조정리2.1]에 의하여 H_1에 의하여 만들어지는 공간은 $\mathcal{L}(\beta_1)$이다.

$$H_2=[\frac{\partial^2 y_2}{\partial X_i \partial X_j}]=h''\beta_1\beta_1^T+h\beta_1\beta_1^T,$$

여기에서 $h'=\partial y_2/\partial y_1$, $h''=\partial^2 y_2/\partial y_1^2$ 이므로, H_2에 의하여 만들어지는 공간도 $\mathcal{L}(\beta_1)$이다.□

정리 4.2 2차미분이 가능한 함수 h와 오차항 δ에 대하여 $X \perp\!\!\!\perp \delta$이고, $y_2=h(y_1)+\delta$ 이면 { $\mathcal{B}_1=\mathcal{B}_2$ }이다.

증명 $B_1=(\beta_1,\cdots,\beta_k)$ 를 H_1에 의하여 만들어지는 부분공간 $\mathcal{B}_1$의 기저라고 할 때,

$$H_2=h''\begin{bmatrix}\frac{\partial y_1}{\partial X_1}\\ \vdots \\ \frac{\partial y_1}{\partial X_p}\end{bmatrix}(\frac{\partial y_1}{\partial X_1},\cdots,\frac{\partial y_1}{\partial X_p})+h'H_1$$

$$=h''[\frac{\partial y_1}{\partial X}][\frac{\partial y_1}{\partial X}]^T+hH_1$$

이다. $U^T=(\frac{\partial y_1}{\partial X_1},\cdots,\frac{\partial y_1}{\partial X_p})^T$ 라고 할때, $H_2=h''UU^T+hH_1$ 이며

$$U=[\frac{\partial y_1}{\partial X}]=g_1\beta_1+g_2\beta_2+\cdots+g_k\beta_k$$

이므로 $U\in\mathcal{L}(\beta_1,\cdots,\beta_k)$ 이다. 따라서 H_2에 의하여 만들어지는 선형부분공간도 H_1에 의하여 만들어지는 선형부분공간과 동일하다. 따라서 $\mathcal{B}_1=\mathcal{B}_2$ 이다.□

다음과 같은 6개의 보조정리를 이용하여 조건 (i) - (v)에 주어진 $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}$ 사이의 관계를 $\mathcal{B}_1$과 $\mathcal{B}_2$의 기저의 추정량을 이용하여 분석할 수 있다.

보조정리 4.1 두 선형부분공간 $\mathcal{B}_1$과 $\mathcal{B}_2$에서 $\mathcal{B}_1\perp\mathcal{B}_2$의 필요충분조건은 임의의 $B_1\in\mathcal{B}_1$, $B_2\in\mathcal{B}_2$에 대하여 $\mathrm{cov}(B_1^TX,\ B_2^TX)=B_1^TB_2=0$ 이다.

보조정리 4.2 $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_k$, $k \leq \min(k_1, k_2)$를 $B_1^T X$와 $B_2^T X$사이의 모정준상관계수라고 할때, $\beta_1 \perp \beta_2$의 필요충분조건은 $\rho_1 = \rho_2 = \cdots = \rho_k = 0$이다.

따라서 $\beta_1 \perp \beta_2$에 대한 검정은 정준상관계수에 대한 검정방법을 이용할 수 있는데, Eaton(1983)에 주어진 우도비검정방법을 이용하여 다음을 증명할 수 있다.

보조정리 4.3 $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_k$를 $B_1^T X$와 $B_2^T X$사이의 표본정준상계수라고 할때, 귀무가설 $\rho_1 = \rho_2 = \cdots = \rho_k = 0$하에서 검정통계량과 그 분포는 다음과 같다.

검정통계량은 $T = \sum_{i=1}^{k} (1-\gamma_i)^2$이며 $T \sim U(n-k_1-1, k_1, k_2)$이다.

여기에서 n은 표본의 수이고, k_1, k_2는 β_1과 β_2의 계수이며, 확률분포 $U(n,m,p)$ 는

$$U(n,m,p) \sim \prod_{i=1}^{m} U_i$$

$$U_i \sim \text{Beta}\left(\frac{n-p+1}{2}, \frac{p}{2}\right)$$

이다.

증명 Eaton(1983)의 proposition(10.11)을 이용하여 증명.□

예를들면 $k_1 = 1$ 인 경우 귀무가설하에서 검정통계량 T의 분포는

$$T \sim \text{Beta}\left(\frac{n-k_2-1}{2}, \frac{k_2}{2}\right)$$

이다.

보조정리 4.4 $\beta_1 \subset \beta_2$의 필요충분조건은 $(I - P_{B_2})X \perp B_1^T X$ 이다. 여기서 P_{B_2}는 부분공간 β_2에의 정사영행렬이다.

보조정리 4.5 $\beta_2 \subset \beta_1$의 필요충분조건은 $(I - P_{B_1})X \perp B_2^T X$ 이다.

보조정리 4.6 $\beta_1 = \beta_2$ 의 필요충분조건은 $(I - P_{B_1})X \perp B_2^T X$ 이고 $(I - P_{B_2})X \perp B_1^T X$ 이다.

보조정리 4.1 과 보조정리 4.2 는 (i)의 검정에 이용할 수 있으며, 보조정리 4.1 - 보조정리 4.6 은 조건 (ii) - (v)의 검정에 이용할 수 있다. 이와같은 각 검정에 있어서 검정의 신뢰성은 β_1 과 β_2 의 기저의 추정량 $\hat{B}_1$ 과 $\hat{B}_2$의 신뢰성에 의존하고 있는데, pHd방법은 모형 g_1 과 g_2에 의존하는 근사적인 방법이므로 검정의 신뢰성에 대하여는 좀 더 깊은 연구가 필요할 것으로 판단된다.

5. 모의 실험

pHd 방법의 신뢰성을 분석해보기 위하여 다음과 같이 모의실험을 실시하였다. 10개의 설명변수 $X_1, \cdots, X_{10}$ 를 각각 독립적으로 표준정규분포로부터 추출하고, 두개의 오차항 ε_1 과 ε_2는 각각 설명변수와 독립으로 $\varepsilon_1 \sim N(0, 0.25)$ $\varepsilon_2 \sim N(0, 0.0625)$ 로 부터 랜덤하게 추출하였다. 표본의 크기는 400으로 하였고, 표본추출은 R-코드(Cook & Weisberg, 1994)를 이용하였다.

[모의실험1] 보조정리 3.1 의 타당성검증을 위한 모의실험으로 다음과 같이 y_2 가 y_1의 함수일때 y_1과 y_2 의 edr 사이의 관계를 알아본다. $Z = X_1 + 2X_2$ 라고 할 때,

$$y_1 = Z^2 + 3Z + \varepsilon_1 \tag{5.1}$$

$$y_{21} = 5y_1 + \varepsilon_2 \tag{5.2}$$

$$y_{22} = y_1^2 + 3y_1 + \varepsilon_2 \tag{5.3}$$

$$y_{23} = y_1^3 + 2y_1^2 + 3y_1 + \varepsilon_2 \tag{5.4}$$

와 같은 모형을 만들었다. 위 모형에서 $B_1 = (1,2,0,\cdots,0)^T$ 이라고 할 때,

$y_1 \perp\!\!\!\perp X \mid B_1^T X$, $y_{21} \perp\!\!\!\perp X \mid B_1^T X$, $y_{22} \perp\!\!\!\perp X \mid B_1^T X$, $y_{23} \perp\!\!\!\perp X \mid B_1^T X$ 임을 알 수 있다. 따라서 $y_1, y_{21}, y_{22}, y_{23}$ 의 모든 경우에 있어서 B_1의 추정량은 동일하여야 한다.

모형(5.1) - 모형(5.4)에 있어서 B_1의 추정량을 각각 $b_1, b_{21}, b_{22}, b_{23}$ 라고 할 때, [표 5.1]에 이 값들이 주어져 있으며, $Z_1 = \beta^T X$, $b_1^T X = h_{11}, b_{21}^T X = h_{12}, b_{22}^T X = h_{13}, b_{23}^T X = h_{14}$ 라고 할 때, Z_1 $h_{11}, h_{12}, h_{13}, h_{14}$ 의 플롯이 [그림 5.1]에주어져있다.

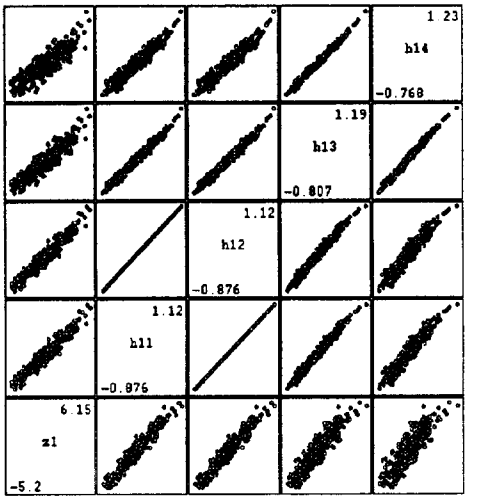
[표 5.1] 과 [그림 5.1]에 의할때, b_1과 b_{21}은 동일하며 따라서 h_{11}과 h_{12}는 완전한 선형관계임을 알 수 있다. 이는 y_{21} 이 y_1의 선형함수이므로 나타난 결과이며 h_{11}과 h_{13} h_{14}로 갈수록 상관성이 줄어드는 것은 y_{22}, y_{23}가 각각 y_1의 2차 3차 함수이기 때문이다. 또한 Z_1 $h_{11}, h_{12}, h_{13}, h_{14}$ 사이의 플롯에 의할때 모선형함수 $\beta^T X$ 와 추정된 선형함수 $b^T X$ 사이에는 선형관계가 있음을 알 수 있다. 한예로 Z_1 의 h_{11}에 대한 최적직선은

$$Z_1 = -0.05 + 6.11\, h_{11} \qquad (\ )\text{안의값은 t 값}$$
$$(-1.635)\quad (69.865)$$
$$R^2 = 0.92$$

으로 h 값이 변환된 것을 고려 할 때 강한 선형관계에 있음을 알 수 있다.

[표 5.1] pHd의 추정벡터

β	b_1	b_{21}	b_{22}	b_{23}
1	0.339	0.339	0.319	0.303
2	0.899	0.899	0.853	0.801
0	-0.001	-0.001	-0.002	0.021
0	0.145	0.145	0.236	0.300
0	-0.022	-0.022	-0.026	-0.031
0	0.050	0.050	-0.191	0.281
0	-0.114	-0.114	-0.171	-0.227
0	-0.058	-0.050	-0.093	-0.121
0	0.144	0.144	0.155	0.144
0	-1.125	-0.125	0.123	-0.096



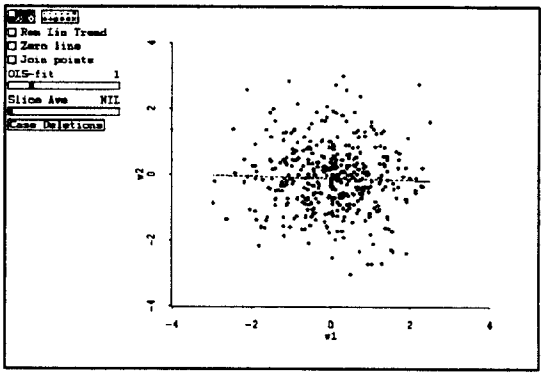
[그림 5.1] h값들의 플롯행렬

[모의실험 2] $Z_1 = X_1 + 2X_2$, $Z_2 = 2X_3 + X_4$ 라고 할 때

$$y_{21} = 2Z_1^2 + Z_1 + \varepsilon_1 \tag{5.5}$$

$$y_{22} = Z_2^2 + 2Z_2 + \varepsilon_2 \tag{5.6}$$

과 같이 정의하였다. 즉 $B_1 = (1,2,0,\cdots,0)^T$, $B_2 = (0,0,2,1,0,\cdots,0)^T$ 라고 할 때, $y_1 \perp\!\!\!\perp X \mid B_1^T X$ 이고 $y_2 \perp\!\!\!\perp X \mid B_2^T X$ 이다. 이 경우 $B_1^T B_2 = 0$ 이므로 B_1과 B_2의 추정량을 각각 b_1과 b_2, $b_1^T X = w_1$, $b_2^T X = w_2$ 라면 w_1과 w_2의 상관계수가 0이어야 한다. w_1과 w_2의 상관계수는 r = 0.034102052로 0에 가까우며 [그림 5.2]의 플롯에 의하여도 w_1과 w_2가 관계가 없음을 알 수 있다.



[그림 5.2] w1 과 w2의 플롯

[모의실험3] [모의실험2]에서 정의된 z_1과 z_2를 이용하여

$$y_{31} = 2Z_1^2 + Z_1 * Z_2 + 3Z_2^2 + 2Z_2 + \varepsilon_1 \quad (5.7)$$

$$y_{32} = Z_2 + 2Z_2 + \varepsilon_2 \quad (5.8)$$

과 같이 정의하였다. 따라서 $B_1 = [\beta_1, \beta_2]$, $B_2 = [\beta_2]$ 라고 정의할 때 $y_1 \perp\!\!\!\perp X \mid B_1^T X$, $y_2 \perp\!\!\!\perp X \mid B_2^T X$ 이며 $\beta_2 \subset \beta_1$이다. pHd방법에 의한 B_1과 B_2의 추정치는 각각

$\hat{B}_1 =$

b_{11}	b_{12}
-0.112	0.235
-0.296	0.855
-0.877	-0.329
-0.323	-0.059
0.060	-0.076
0.030	0.123
-0.009	-0.253
-0.102	-0.014
-0.108	0.133
-0.008	0.000

$\hat{B}_2 =$

b_2
0.049
0.079
-0.915
-0.436
0.038
0.101
-0.115
-0.098
-0.012
0.029

와 같다. $b_{11}^T X = w_{31}$, $b_{12}^T X = w_{32}$ $b_2^T X = w_2$라고 할 때, $\beta_2 \subset \beta_1$이므로 w_2는 w_{31}과 w_{32}의 선형결합에 의하여 표현될 수 있어야 함을 예측할 수 있다. w_2와 $[w_{31}, w_{32}]$의 3차원 플롯이 [그림 5.3]에 주어져 있고, w_2의 w_{31}과 w_{32}에 대한 회귀선은

$$w_2 = \underset{(218.5)}{0.889 w_{31}} + \underset{(101.3)}{0.423 w_{32}} \qquad (\ ) \text{ 안의 값은 t값}$$

$$R^2 = 0.993203$$

으로 예측한 결과가 나타났다고 할 수 있다. 또한 [보조정리4.4]에 의하여 $(I - P_{B_1})^T X$ 과 $B_2^T X$가 서로 직교하므로 $(I - P_{B_1})X$ 와 $B_2^T X$사이에 정준상관계수가 0이어야 한다. 여기에서 B_2의 계수가 1이므로 정준상관계수의 추정량은 $w_2 = \hat{B}_2^T X$에 대한 $(I - \hat{P}_{B_1})X$ 의 회귀식에 의하여 구하여지는 다중상관계수와 동일하다. $w_2$hat를 w_2의 $(I - \hat{P}_{B_2})X$ 에 대한 회귀식의 추정량이라고 할 때 [그림5.4]에 w_2와 $w_2$hat사이에 플롯이 주어져 있는데, 이에 의할 때 w_2의 $(I - \hat{P}_{B_2})X$ 에대한 다중상관계수가 r = 0.1520(R^2 = 0.0231) 으로 거의 0에 가까움을 알 수 있다.

6. 결 론

5절에서 주어진 모의 실험에 의할때, pHd방법이 edr공간을 어느정도 신뢰할 수 있을만큼 추정가능하다는 것을 알 수 있다. 그러나 앞의 모의실험에서는 pHd방법이 가장 잘 적합되는 2차식에 근거하여 차수를 높여나갔다. 특히 [모의실험1]에서 차수가 증가 될 수록 edr공간의 추정이 부정확해짐은 주의하여 관찰하여야 할 사항이다. 이와같은 다변량반응 모형에대한 변수차원축약방법은 변수더하기 플롯(added variable plot)과 직접적인 관계가 있으며 부분잔차플롯(partial residual plot)등과 연결하여 이용할 수 있다.

참 고 문 헌

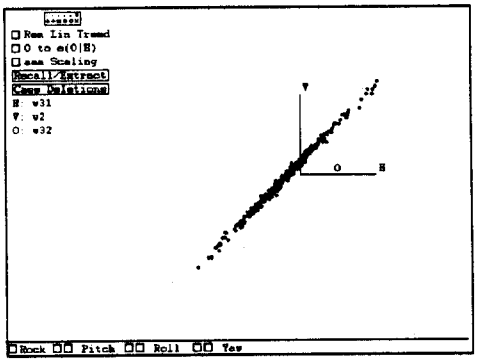
[1] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression" by K.C. Li, *Journal of the American Statistical Association*, Vol. 86, 328-332.

[2] ———————— (1994). *Introduction to Regression Graphics*, unpublished manuscript, University of Minnestoa.

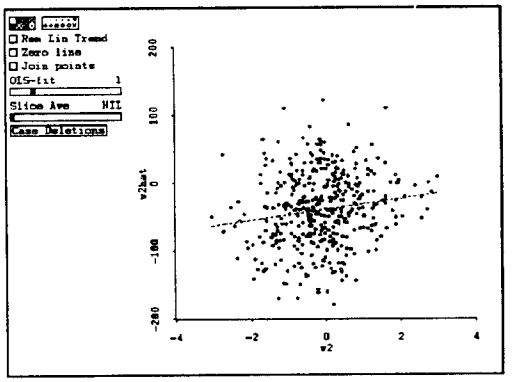
[3] Eaton, M. L. (1993). *Multivariate Statistics*, A Vector Space Approcah, Wiley.

[4] Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, Vol. 86, 316-342.

[5] ———— (1992). On Principal Hessian Directions for data visualization and dimension reduction : Another application of Stein's Lemma, *Journal of the American Statistical Association*, Vol. 87, 1025-1040.



[그림 6.3] w2와 [w31, w32]의 플롯



[그림 6.4] w2와 w2hat의 플롯

Multivariate pHd Analysis

Yong Goo Lee[3)]

Abstract

These days, many kinds of graphical methods have been developed, and it is possible to get information directly from data. Especially, R-code (Cook and Weisberg ;1994) make it possible to draw various kinds of two and three dimensional plots, and to rotate the axis of the plots. But the maximum dimension of the plot is three, so we can not draw plot of one response variable with more than three explanatory variables. Li(1991,1992) has developed a method to reduce the dimension of the explanatory variables, so it is possible to draw lower dimensional plots to get information of the full explanatory variables. One of the dimension reduction method developed by Li is pHd. In this paper, we have tried to apply the pHd method for the model with multivariate response.

3) Dept of applied statistics, ChungAng University, HeuckSuk Dong 221, DongJakGu, Seoul 156-756, KOREA.