제2장

차원축소기법을 이용한 무응답 처리방안

최 필 근

제1절 서 론

무응답은 대부분의 통계조사에서 흔히 발생하며, 현실적으로 무응답이 발생하는 것을 피하기는 어려운 일이다. 이러한 무응답은 조사 결과에 상당한 영향을 주기 때문에 조사의 신뢰도를 높이기 위해서는 무응답률을 낮추는 일이 매우 중요하다. 하지만 많은 노력에도 불구하고 시간·공간적인 여건과 비용의 제약으로 인하여 무응답은 불가피하게 발생하며, 특히 최근에는 사생활보호에 의한 답변 기피로 인하여 무응답이 늘어가고 있는 실정이다. 따라서 이러한 무응답에 대해서 사후적으로 처리하는 방법을 고려해야 한다.

무응답의 발생 원인은 크게 두 가지로 나눌 수 있다. 응답자를 만나지 못하거나 응답자가 응답을 거절하는 경우에 질문 문항에 대하여 전체적인 무응답이 발생하게 된다. 이러한 경우에 발생하는 무응답을 단위 무응답(unit nonresponse)이라고 한다. 그리고 부분적인 문항에 대하여 응답을 받지 못해서 생기는 무응답이 있을 수 있는데, 이러한 경우의 무응답을 항목 무응답(item nonresponse)이라고 한다. 많은 경우 무응답자의 성향은 응답자의 성향과 다르므로 무응답을 무시하고 응답자만으로 모집단을 추정할 경우 편향(bias)이 발생하며 이 때 발생하는 편향을 무응답 편향(nonresponse bias)이라 한다. 이러한 무응답 편향을 줄이기 위

해 여러 가지 보정 방법들이 연구되었는데, 크게 단위 무응답에 대한 보정 방법으로는 일반적으로 무응답 가중치 조정 방법이 널리 이용되고 있으며, 항목 무응답에는 무응답 대체(imputation)방법이 주로 이용되고 있다.

본 연구에서는 차원축소기법을 이용하여 항목 무응답을 대체하기 위한 새로운 방법을 제시하고자 한다. 항목 무응답 대체방법은 여러 가지가 있으며, 개별 문제에 따라 적절한 대체 방법을 채택하여 이용하는 것이 통상적인 방법이다. 대체방법의 사용목적은 결측값을 다른 값으로 대체한 완전한 자료를 구성하여 이에 따라 기존의 완전한 자료에 적용되는 통계분석기법을 그대로 적용할 수 있게 하는 것이다. 이때 실제 결측값과 대체값의 차이가 적을수록 신뢰도가 높은 조사 결과를 이끌어낼 수 있을 것이다. 이러한 결과를 이끌어내기 위해서는 무엇보다도 목표변수의 값들을 잘 대체하기 위한 보조변수의 활용이 중요하다고 할 수 있다. 많은 보조변수로부터 목표변수를 추정하기 위한 정보를 이끌어낸다면 더 좋은 대체, 즉 대체후 분산과 편향을 줄일 수 있게 될 것이다.

현실적으로 추정을 위한 모형을 구축하는 데 있어 많은 보조정보를 사용하지 못한다. 이는 많은 보조변수로부터 오는 모형의 복잡성으로 추정의 정도를 오히려 감소시킬 수 있기 때문이다. 다시 말해 차원의 증가로 인해 모형의 구축이 힘들어지는 것이다. 따라서 일반적으로 목표변수와 상관도가 높은 변수 또는 영향을 많이 줄 것 같은 변수들을 선택하여 보조정보로 이용을 한다. 그러나 이러한 경우 저차원으로 모형의 복잡성은 해결되나 포기하는 보조변수가 많아질수록 실제 모형을 추정함에 있어서 정보의 손실로 인해 추정의 정도는 낮아질 수밖에 없다. 따라서 이러한 문제를 해결하기 위하여 차원축소기법을 이용하여 무응답을 대체하는 방법에 적용하고자 한다.

차원축소기법은 보조변수의 정보손실이 없이 고차원을 저차원으로 만들어 모형을 추정하는 새로운 방법으로 무응답 대체에 적용시키기 위하여 대체시 보조변수를 이용하는 대체방법인 회귀 대체방법(regression imputation method)과 최근방이웃 대체방법(nearest-neighbor imputation method)을 고려하였다.

본 연구의 주요 목적은 차원축소기법을 이용하여 항목 무응답을 대

체하는 방법을 제시하고, 기존의 방법들과의 비교를 통하여 제시된 방법의 효율성을 보이고자 함에 있다. 이를 위하여 제2절에서는 무응답을 대체하는 여러 가지 방법을 설명한다. 제3절에서는 차원축소의 정의와 필요성을 설명하고 차원축소를 위한 새로운 방법인 커널 등고선회귀(kernel contour regression)방법을 소개한다. 제4절에서는 모의실험을 통하여 기존의 평균대체, 회귀대체 및 최근방대체 방법과 본 연구에서 제시된 커널 등고선회귀기법을 이용한 방법을 비교·분석하고자 한다. 마지막으로, 제5절에서는 본 연구의 최종적인 결론과 더불어 향후 연구되어야 할 사항들을 제시하고자 한다.

제2절 무응답 대체방법

이 절에서는 주로 많이 쓰이고 있는 무응답 대체방법(Imputation Method)과 그 내용에 대하여 간략히 소개하고자 한다.

1. 개요

통계조사에서 자주 발생하는 무응답은 조사의 정도(precision)를 저하시키므로 적절한 무응답 대체방법에 의하여 결측값을 다른 값으로 대체하여 조사의 정도를 높일 필요가 있다. 대체방법으로는 결측값의 대체값으로 한 개의 값을 부여하는 단일대체법(single imputation)과 여러 개의 값을 대체하는 다중대체법(multiple imputation, Rubin(1987))으로 구분된다. 다시 세부적으로 단일대체법은 결측된 값에 유일하게 결정된 대체값을 대입을 하는 결정적 대체법(deterministic imputation)과 대체값을 확률적으로 결정하여 대입을 하는 확률적 대체법(stochastic imputation)으로 나누어진다. 이들 중 몇 가지 주요 방법들을 소개한다.

2. 결정적 대체법

평균대체(mean imputation)는 결측자료에 대하여 목표변수의 전체 평

균을 대입하거나 몇 개의 대체군(imputation class)으로 분류한 후 각 층에서의 응답자 평균값으로 대체하는 방법이다. 즉,

$$y_k^* = \begin{cases} y_k\,, & k \in R, \quad \text{응답값} \\ \overline{y_r}, & k \in R^c, \quad \text{대체값} \end{cases}$$

여기서 R은 응답 집합, R^c은 무응답 집합이며 $\overline{y_r}$은 응답값의 평균이다. 이 방법은 사용이 매우 쉬운 장점이 있으나, 항목변수가 양적변수이고 구하고자 하는 통계량이 평균일 때 유용하다. 그러나 대체 후의 값들은 평균값의 빈도수가 지나치게 많아져 응답값들의 분포가 왜곡될 가능성이 있다.

최근방대체(nearest-neighbor imputation)는 결측자료가 발생하는 목표변수와 이 목표변수에 대한 보조변수와의 거리가 가장 유사한 응답값으로 결측자료를 대체하는 방법이다. 즉,

$$y_k^* = \begin{cases} y_k\,, & k \in R, \quad \text{응답값} \\ y_i,\ i : \min_{i \in R} \| x_k - x_i \|\,, & k \in R^c, \quad \text{대체값} \end{cases}$$

여기서 거리가 가장 가까운 응답값을 찾는 방법은 여러 가지가 있으며, 주로 절대거리를 많이 이용한다. 이 방법은 활용가능성은 높지만 주어진 보조변수가 적절치 않거나 무응답률이 높은 경우 추정에 큰 편향을 가져올 수 있다.

회귀대체(regression imputation)는 결측자료가 발생하는 목표변수(y)와 이 목표변수에 대한 보조변수($x_1, x_2, \cdots, x_k$)로 회귀모형을 적합시킨 후에 추정값에 의하여 대체를 하는 방법이다. 즉,

$$y_k^* = \begin{cases} y_k\,, & k \in R, \quad \text{응답값} \\ \hat{y_r}, & k \in R^c, \quad \text{대체값} \end{cases}$$

여기서 $\hat{y_r}$은 응답값과 그에 따른 보조변수를 사용해서 최소제곱방법에 의하여 회귀계수를 추정한 후 결측값의 보조변수 값을 대입하여 구한 값이다. Greenless et al.(1982)은 미국의 인구조사(CPS: current population survey)에서 발생하는 결측값을 대체하기 위해서 회귀대체 방법을 이용

하였으며, 그 결과 대체된 값과 실제값의 평균절대편차(mean absolute deviation)를 비교할 때 다른 대체 방법에 비해 매우 적절함을 보였다.

이 방법들 외에도 축차 핫덱대체(sequential hot-deck imputation)는 대체군 내에 응답자료를 순서대로 정리하여 결측값이 있는 경우 그 결측값 바로 이전의 응답을 결측값 대신 대체하는 방법이다. 논리적대체(logical imputation)는 논리적인 제약조건이나 다른 기록에 의하여 확실하게 대체값을 지정하여 결측값을 대체하는 방법이다. 주로 항목간에 분명한 논리적인 관계가 있을 때 적용이 가능하다. 시기적대체(historical imputation)는 조사시점과 조사항목 간의 높은 상관관계가 있을 때 조사시점의 결측값을 과거시점의 조사값으로 대체하는 방법이다.

3. 확률적 대체법

결정적 대체법은 대체값을 유일하게 결정하기 때문에 목표변수의 변동을 줄이는 경향이 있다. 이러한 단점을 보완하기 위하여 확률적 대체법이 제안되었으며, 이 방법은 대체값에 확률적인 변동을 부여해 줌으로써 결정적 대체법의 단점을 보완할 수 있다.

핫덱대체(hot-deck imputation)는 대체군 내의 응답값 중에서 하나를 임의로 선택하여 대체하는 방법으로 대체 후에도 표본의 분포가 그대로 유지될 수 있다는 장점이 있다. 그러나 응답 패턴이 목표변수와 무관한 경우에 적합하다. 가중핫덱대체(weighted hot-deck imputation)는 응답값 중 하나를 선정할 때 가중값을 두어 선정하는 방법이다. 이 방법은 층화나 집락화가 되어 있어 대체군 내의 응답값들이 서로 다른 추출 확률이 될 때 핫덱대체 대신에 이용하는 방법이다. 랜덤회귀대체(random regression imputation)는 회귀대체값에 확률오차를 포함시켜 대체하는 방법이다. 확률적 대체법은 결측값 대체 전과 대체 후의 변동이 유지되도록 하는 장점이 있는 반면 대체하는 과정이 다소 복잡하고 계산이 많다는 단점이 있다.

제3절 차원축소기법

이 절에서는 차원축소에 관하여 알아보고 새로운 차원축소기법(Dimension Reduction)인 커널 등고선회귀 방법에 대하여 알아보고자 한다.

1. 차원축소의 정의와 필요성

차원축소에 관한 이론은 1990년대 초반부터 Cook(1992, 1994, 1998) 등에 의하여 다양한 연구논문들이 발표되기 시작하였다. 이러한 이론의 주요 목적은 고차원에서 나타나는 많은 문제점들을 차원축소를 통하여 해결하고자 하는 데 있다.

회귀모형에서 차원축소란 회귀모형이 본질적으로 내포하고 있는 정보의 손실 없이 고차원에서 저차원으로 축소시킨 공간을 찾는 것을 말한다. 차원축소의 의미를 쉽게 풀어서 설명한다면 다음과 같이 말할 수 있을 것이다.

> "하나의 쓰레기통이 있다고 하자. 그리고 채워진 쓰레기통의 부피를 차원, 쓰레기를 모형을 추정하기 위한 정보라고 하자. 우리는 정보를 많이 이용하기 위하여 쓰레기를 쓰레기통 안에 모을 것이다. 많은 정보를 모을수록 쓰레기통은 꽉 채워지고 차원은 높아져 추정이 어렵게 될 것이다. 이 때 차원축소를 한다. 즉, 발로 쓰레기통을 힘껏 누를 것이다. 그러면 우리가 정보라고 생각한 것은 통 속에서 조금도 빠져나가지 않을 것이다. 그러나 쓰레기통의 부피는 상당히 줄어들어 있을 것이다. 똑같은 정보하에 차원만 줄어든 것이다. 이것이 바로 차원축소의 본질인 것이다."

이러한 내용을 수리적인 측면에서 살펴보면 다음과 같은 식으로 이루어짐을 알 수 있다. 스칼라(scalar)인 반응변수(목표변수) y와 p개의 설명변수(보조변수)들로 이루어진 $p\times 1$벡터 $X=(X_1,X_2,\cdots,X_p)\in R^p$를 고려하자. 이 때 차원축소의 궁극적인 목표는 다음과 같이 표현할 수 있다.

$$y\perp X|\eta^T X \qquad (1)$$

여기서 η는 $p\times q\,(q\le p)$행렬이며 $\perp$는 독립의 의미로 쓴 것이다. 식 (1)은 설명변수 X의 모든 값에 대하여 $\eta^T X$가 주어졌을 때 y와 X가 독립이라는 의미로 $y|X$의 분포는 $y|\eta^T X$의 분포와 같음을 의미한다. 즉, $F(\bullet)$을 누적분포(cumulative distribution)라 할 때 $F(y|X)=F(y|\eta^T X)$로 표현할 수 있다. 따라서 만일 $p\times q$ 행렬 η를 발견할 수 있다면 p차원인 원래변수 X를 p보다 작은 q차원의 새로운 변수 $\eta^T X$로 대체할 수 있고 이 때 $y|X$의 조건부 확률이 $y|\eta^T X$의 조건부 확률과 같아져서 X 대신 $\eta^T X$로 회귀모형을 설명할 수 있게 된다. 그러므로 회귀모형의 정보 손실이 없이 p차원의 설명변수를 p보다 작은 q차원으로 축소시켜 모형을 구축할 수 있다.

저차원의 설명변수를 만들기 위한 핵심과정은 식 (1)의 η를 찾는 데 있다. 여기서 $p\times q$ 행렬 η의 열벡터로 이루어진 부공간 $S(\eta)$를 차원축소 부공간(dimension reduction subspace)이라고 한다. 차원축소 부공간은 유일하지 않고 많이 존재하게 되며 이러한 부공간들 중에서 가장 효율적인 최소 차원축소 부공간을 찾아야 한다. 따라서 존재하는 모든 가능한 차원축소 부공간에 대하여 교집합인 $\cap S(\eta)$을 구함으로써 가장 최소한의 부공간을 구할 수 있다. 이러한 부공간이 최종적으로 추정해야 할 차원축소 부공간이며 이를 중심 부공간(central subspace, $S_{y|X}$)이라고 한다.

이러한 중심 부공간을 추정함으로써, 많은 보조변수의 활용으로 생기는 고차원의 문제점들을 정보의 손실이 없이 저차원의 관점에서 해결이 가능해진다. 그러므로 주어진 정보를 최대한 활용하여 무응답 대체 방법에 적용함으로써 대체의 정도를 높일 수가 있을 것이다. 중심 부공간을 추정하는 방법은 1990년대 초반 Cook, Weisberg(1991)와 Li(1991, 1992) 등에 의하여 연구되기 시작하였으며, 본 연구에서는 여러 분야에서 기존의 중심 부공간을 추정하는 방법보다 더 효율적인 방법인 커널 등고선회귀(최 · 이, 2005) 방법을 무응답 대체를 위해 이용하고자 한다. 다음 항에서 커널 등고선회귀 방법을 소개한다.

2. 커널 등고선회귀 방법

커널 등고선회귀(Kernel Contour Regression) 방법은 반응값(y)들의 차이가 작게 형성되는 관찰값(x)들의 차이벡터를 추출하여 자료로 이용을 하는데, 이러한 자료의 집합을 등고선방향들(contour directions)이라 하며 이 방향들은 추정하고자 하는 중심 부공간의 방향을 형성한다. 따라서 등고선 방향을 가지고 등고선회귀를 구축하면 중심 부공간을 찾을 수 있다.

가. 등고선회귀

(X^*, y^*)를 (X, y)의 독립적인 표본이라고 하자. 이 때 중심 부공간에 속하는 벡터를 $v \in S_{y|X}$라 하고 중심 부공간에 직각인 벡터를 $w \in (S_{y|X})^{\perp}$라 할 때 상수 $c > 0$에 대하여 다음의 가정 사항을 가진다.

$$\mathrm{Var}\left[w^T(X^* - X) \,|\, |y^* - y| \le c\right] > \mathrm{Var}\left[v^T(X^* - X) \,|\, |y^* - y| \le c\right]. \quad (2)$$

식 (2)의 가정은 실제 $y|X$의 조건부 분포가 w^TX에는 의존하지 않고 v^TX에만 의존하기 때문에 반응값들의 변동은 w^TX에서보다도 v^TX에서 훨씬 크게 된다. 그렇기 때문에 반응값의 같은 증분 안에서는 반대로 w^TX에서 크게 됨을 직관적으로 알 수 있다. 이 가정에 관한 사항은 여러 가지 경우에 대해서 모의실험을 한 결과 위배되는 경우를 보이지 않았다(최, 2006). B. Li(2005)는 앞의 가정 사항하에서 등고선회귀를 위한 행렬을 다음의 식 (3)과 같이 나타내었다.

$$K(c) = \mathrm{E}\left[(X^* - X)(X^* - X)^T \,|\, |y^* - y| \le c\right]. \quad (3)$$

이 때 실제 중심 부공간의 차원이 q라면 가장 작은 q개의 고유값(eigenvalue)에 해당하는 행렬 $K(c)$의 고유벡터(eigenvector)들이 중심 부공간을 생성하게 됨을 보였다. 이 결과를 토대로 커널 등고선회귀 방법에 의하여 중심 부공간을 추정한다.

나. 커널 등고선회귀

1) 등고선회귀의 문제점

앞에서 언급한 등고선회귀는 다음의 문제점을 가지고 있다. 첫째, 응답값들의 차이가 크고 작음에 관계없이 추출된 등고선 방향들은 똑같은 정보를 가지고 있다는 가정하에서 추정을 한다. 그러나 실제로는 반응값들의 차이가 작을수록 등고선 방향들은 더 많은 정보를 가지고 있는 것이다. 둘째, 반응값 차이를 정하는 c값이 변함에 따라 추출되는 등고선 방향들도 계속 변하게 된다. 따라서 최종적으로 추정하고자 하는 중심 부공간 역시 계속적으로 변하여 추정된 중심 부공간은 일정(robuss)하지 못하게 된다. 셋째, 반응값 차이를 정하는 c값에 대해서 가장 적절한 값을 정하기란 현실적으로 불가능하다고 할 수 있다.

이러한 문제점을 해결하기 위하여 커널을 이용하여 등고선회귀를 구축한 새로운 방법을 제시하였다(최, 2006).

2) 커널함수(Kernel Function)의 개요

커널함수는 국소이웃(local neighborhood)의 본성을 특정화하여 회귀함수의 추정에 제공되는 가중치라고 할 수 있다. 즉, 추정에 사용되는 값과 가깝게 위치할수록 많은 가중치를 주어서 실제 추정값에 많이 반영하는 것이다. 회귀함수의 커널추정방법들은 1960년대부터 많이 연구되기 시작하였다. 초기 대표적인 추정량은 Nadaraga-Watson(1964)에 의해서 제안되었으며 그 형태는 식 (4)와 같다.

$$\hat{f}(X_0) = \frac{\sum_{i=1}^{n} K_\lambda(X_0, X_i) y_i}{\sum_{i=1}^{n} K_\lambda(X_0, X_i)} \tag{4}$$

여기서 $K(\bullet)$는 대칭인 임의의 확률밀도함수로서 커널이라고 불리고, λ는 평활량(bandwidth)을 좌우하는 평활계수로서 표본의 수가 커질수록 0으로 가까이 가는 성질을 만족한다. 일반적으로 $f(X_0)$의 국소 회귀추정량은 추정모수 θ에 대하여 식 (5)를 최소화함으로써 얻을 수 있다.

$$\sum_{i=1}^{n} K_\lambda(X_0, X_i)(y_i - f_\theta(X_i))^2 \qquad (5)$$

여기서 함수 $f_\theta(X)$가 상수함수(constant function)이면 식 (4)와 같은 Nadaraga-Watson 추정량의 형태로 나오게 되며, 이를 이용하여 커널 등고선회귀를 유도하고자 한다.

3) 커널 등고선회귀의 표본 추정량

식 (5)에서 커널함수 $K_\lambda(X_0, X_i)$을 대신하여 $K_\lambda(0, y_j - y_i)$을 사용함으로써 반응값의 차이가 작게 되는 등고선 방향들에 대해서 많은 가중치를 할당하게 되는 것이며, $(y_i - f_\theta(X_i))$ 대신 $(U(X_i, X_j; c) - A$ (constant function))을 사용하여 추정량을 구하게 되는 것이다.

여기서 $U(X_i, X_j; c) = (X_i - X_j)(X_i - X_j)^T I(|y_j - y_i| \le c)$은 식 (3)에서 나왔으며 이것은 상수함수로 추정을 하게 된다. 따라서 다음의 식 (6)을 최소화함으로써 커널 등고선회귀의 표본 추정량을 구할 수 있다.

$$\sum_{(i,j)\in N} K_\lambda(0, y_j - y_i)(U(X_i, X_j; c) - A(\text{constant function}))^2 \qquad (6)$$

N은 $(i,j)\colon i = 2, \cdots, n; j = 1, \cdots, i-1$의 집합을 의미한다. 최종 추정량은 앞에서 언급한 Nadaraga-Watson 추정량의 형태로 유도됨을 알 수 있다.

$$\widehat{KCR}(c) = \frac{N_c}{C_2^n} \frac{\sum_{(i,j)\in N \ |y_j - y_i| \le c} K_\lambda(0, y_j - y_i)(X_i - X_j)(X_i - X_j)^T}{\sum_{(i,j)\in N \ |y_j - y_i| \le c} K_\lambda(0, y_j - y_i)}$$

$$K_\lambda(0, y_j - y_i) = \frac{1}{\lambda}\left[-\frac{(y_j - y_i)^2}{2\lambda}\right] \qquad (7)$$

식 (7)에서 사용할 커널함수는 여러 가지가 있으나, 본 연구에서는 가장 일반적이고 모든 등고선 방향에 대하여 가중치가 적용되는 Gaussian 커널함수를 사용하였으며, N_c는 $|y_j - y_i| \le c$을 만족하는 개수이다. 식 (7)에서의 추정량은 앞에서 언급한 등고선회귀의 문제점을 보완해 준다. 즉, 반응값들의 차이가 작을수록 등고선 방향들에 대하여 더 많은 가중

치를 주고 있으며, 반응값 차이를 정하는 c값에 관계없이 추출된 등고선 방향에 대하여 적절하게 가중치를 부여함으로써 항상 일정한 중심 부공간을 추정할 수 있도록 해준다.

다. 중심 부공간 추정

커널 등고선회귀 방법에 의하여 본 연구의 핵심이라고 말할 수 있는 중심 부공간을 추정하는 절차는 다음과 같다.

step 1 : 설명변수(보조변수, X)의 표본평균 벡터와 표본분산 행렬을 구한다.

$$\hat{\mu}= \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad \hat{\Sigma}= \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})(X_i-\hat{\mu})^T$$

step 2 : 식 (7)에서 커널 등고선회귀의 표본 추정량 $\widehat{KCR}(c)$을 구한다.

step 3 : 행렬 $\widehat{\Sigma^{-1/2}}\widehat{KCR}(c)\widehat{\Sigma^{-1/2}}$의 특이값 분해를 한다. 만일 표준화된 설명변수 Z의 실제 중심 부공간의 차원이 q라면 가장 작은 q개의 고유값에 상응하는 고유벡터 $\widehat{\eta_{p-q+1}},\cdots,\widehat{\eta_p}$를 구한다.

step 4 : 이 고유벡터는 중심 부공간 $S_{y|Z}$의 추정량을 생성한다. 따라서 처음 설명변수의 중심 부공간 $S_{y|X}$을 추정하기 위해서 $\widehat{\Sigma^{-1/2}}$을 고유벡터 $\widehat{\eta_{p-q+1}},\cdots,\widehat{\eta_p}$에 곱하여 $\widehat{\Sigma^{-1/2}}\widehat{\eta_{p-q+1}},\cdots,\widehat{\Sigma^{-1/2}}\widehat{\eta_p}$을 구한다. 이 벡터가 최종적으로 추정하고자 하는 중심 부공간이 된다.

3. 커널 등고선회귀의 차원 결정

커널 등고선회귀 방법으로 중심 부공간을 추정하는 것을 보았다. 실제로는 이보다 앞서 차원을 결정하여야 한다. B. Li(2005)는 차원을 결정하는 방법으로 식 (2)의 가정 사항과 설명변수의 정규성 조건이 맞는다면 다음과 같이 결정할 수 있음을 보였다.

실제 중심 부공간의 차원이 q라면 $\Sigma^{-1/2}K(c)\Sigma^{-1/2}$의 $p-q$개의 큰 고유값은 2가 된다는 것이다. 이는 중심 부공간의 직교 방향의 측면에서 고려한 것으로 중심 부공간의 측면으로 고려한다면 $2I_p - \Sigma^{-1/2}K(c)\Sigma^{-1/2}$의 q개의 큰 고유값이 2가 된다는 의미와 같다. 그리고 $p-q$개의 작은 고유값은 0이 될 것이다.

본 연구에서는 이 방법을 커널 등고선회귀에 적용하여 차원의 추정을 할 것이다. 표본을 가지고 구한 $2I_p - \widehat{\Sigma^{-1/2}}\widehat{KCR}(c)\widehat{\Sigma^{-1/2}}$의 고유값을 구하여 1보다 큰 고유값에 대하여 중심 부공간 추정을 위한 차원으로 인정을 할 것이다.

여기서 중심 부공간 추정과 다른 점은 모든 가능한 등고선 방향을 사용하는 것이 아니라 추출된 등고선 방향들에 대하여 차원을 추정하는 것이기 때문에 식 (7)에서 C_2^n이 $|y_j - y_i| \le c$을 만족하는 개수(N_c)로 바뀐 것이다.

즉, 다음의 $\widehat{KCR}(c)$을 사용하여 차원을 추정한다.

$$\widehat{KCR}(c) = \frac{\sum_{(i,j)\in N\ |y_j-y_i|\le c} K_\lambda(0, y_j - y_i)(X_i - X_j)(X_i - X_j)^T}{\sum_{(i,j)\in N\ |y_j-y_i|\le c} K_\lambda(0, y_j - y_i)}$$

커널 등고선회귀의 차원 추정에 관한 검정통계량은 연구되어야 할 과제 중의 하나이다. 이 방법의 대안으로 비모수적으로 접근한 순열검정(permutation test)을 이용할 수도 있다(Cook & Yin, 2001).

제4절 모의실험

항목 무응답 대체를 위한 새로운 방법으로 차원축소기법인 커널 등고선회귀를 이용한 방법을 설명하였다. 새로이 제시된 방법의 효율성을 알아보기 위하여 보조변수들을 이용하여 무응답을 대체하는 방법들 중 주로 많이 쓰이는 회귀대체 방법과 최근방대체 방법의 비교・분석을 실시하였다. 무응답 대체를 위해 많이 사용되는 핫덱 방법 등은 직접적으로 보조변수들을 활용하지 않기 때문에 새로운 방법과 비교하기는 힘들

지만, 회귀대체 방법과 최근방대체 방법의 비교·분석 연구가 많다(김규성, 2000; 김진, 2004 등). 따라서 핫덱 방법과의 비교는 본 모의실험을 통해서 간접적으로 가능할 것으로 본다.

1. 모의실험 개요

모의실험을 위해서 보조변수는 10개를 생성하였으며 생성을 위한 분포는 정규분포와 감마분포 두 가지를 고려하였다. 감마분포의 사용은 보조변수의 왜도(skewness)를 고려한 실험으로 커널 등고선회귀 방법은 보조변수가 정규분포 이외에서도 다른 차원축소 방법에 비해 추정의 우수성을 보였다(최, 2006). 정규분포인 경우 보조변수는 평균과 분산의 값을 다양하게 생성하였다. 이는 실제 보조변수의 다양성을 고려한 것이다. 감마분포인 경우 적당한 왜도가 나오게 보조변수를 생성하였다. 오차 ϵ는 정규분포에서 독립적으로 생성하였고, 표본의 크기는 $n=300$이다.

목표변수는 보조변수와 오차를 이용하여 다음의 모형으로부터 생성하였다.

$$y=(X_1-X_3+X_4-X_5+X_6-X_9+X_{10})^S+\epsilon \qquad (8)$$

식 (8)에서 목표변수를 생성하기 위하여 보조변수는 7개를 사용하였다. 위에서 제시된 모형은 특별한 의미가 있는 것은 아니며, 실제의 모형이 이 식과 같다면 이러한 식을 어떻게 잘 추정하는가를 보여주기 위함이다. 실제로 목표변수를 추정함에 있어 7개의 보조변수로부터 모든 정보를 이끌어내고 모형의 형태를 알 수 있다면 매우 정확한 추정을 할 수 있을 것이다. 모형의 형태는 1차 선형과 2, 3차 곡선 형태 3가지 ($S=1,2,3$)를 고려하였다. 무응답이 일어나는 형태는 크게 2가지를 고려하였다. 첫 번째는 무응답이 목표변수와 무관하게 일어나는 것을 가정하고 정규분포하에서 생성된 목표변수의 값에 대해서 10%, 20%, 30%, 40%, 50%의 결측치를 랜덤하게 발생시켰으며, 두 번째는 무응답이 목표변수의 큰 값에서 일어나는 것을 가정하여 감마분포하에서 생성

된 목표변수의 중위수 이상의 값에 대해서만 3%, 5%, 7%, 10%, 15%의 결측치를 랜덤하게 발생시켰다. 그리고 각각의 경우에 대해서 $R=1000$ 번을 반복하여 실험하였다.

목표변수의 평균추정에 대한 편향정도를 알아보기 위해 결측 전의 평균값과 5가지 방법에 의해 대체된 후의 추정 평균값의 평균절대오차(mean absolute error)를 계산하고, 추정량의 안정성을 알아보기 위해 추정량의 표준편차를 평균값으로 나누어준 변동계수(CV)를 계산하여 대체방법들의 효율성을 비교하였다.

5가지 방법은 평균대체(M), 회귀대체(R), 커널 등고선회귀를 이용한 회귀대체(KCRR), 최근방대체(N), 그리고 커널 등고선회귀를 이용한 최근방대체(KCRN)이며 평균절대오차의 식은 다음과 같다.

$$MAE_{\mathrm{mean}} = \frac{1}{R}\sum_{i=1}^{R} |M - \widehat{M}_i|$$

여기서 R은 반복횟수를 나타내며 M은 목표변수의 실제 평균이고 $\widehat{M}_i$는 i번째 반복에서 모든 결측값을 각 추정방법에 의해서 대체한 후의 목표변수 추정 평균이다. 모의실험의 내용을 정리하면 다음과 같다.

보조변수 생성분포	모형(S)	무응답률	대체방법
정규분포	1 2 3	10%, 20%, 30%, 40%, 50%	M, R, KCRR, N, KCRN
감마분포	1 2 3	3%, 5%, 7%, 10%, 15%	M, R, KCRR, N, KCRN

2. 모의실험 내용 및 분석결과

가. 정규분포로 보조변수 생성

1) $y=(X_1-X_3+X_4-X_5+X_6-X_9+X_{10})^2+\epsilon$ **모형 (모의실험 I)**

이 모형에 의해 생성된 목표변수와 10개의 보조변수를 가지고 회귀대체와 최근방대체를 하기 위한 보조변수를 선택하였다. 상관분석 및 변수선택 방법에 의한 결과 보조변수 X_1, X_3, X_4, X_5, X_{10}을 회귀대체를 위한 모형 추정에 사용하는 것이 가장 적절하다고 판단되며, 목표변수와 가장 상관도가 높은 X_5는 최근방대체를 위한 보조변수로 사용하였다(보조변수 선택을 위한 결과는 부록을 참조). 따라서 결측된 값은 다음의 추정량으로 대체가 되었다.

$$\hat{y}=\hat{\beta_0}+\hat{\beta_1}X_1+\hat{\beta_2}X_3+\hat{\beta_3}X_4+\hat{\beta_4}X_5+\hat{\beta_5}X_{10}$$

여기서 회귀분석의 결과 선형모형으로 잘 설명될 수 있음을 알 수 있다. 그러나 실제의 모형이 2차의 형태를 가진다는 것을 우리는 알 수 없으며, 이것은 4차원 이상은 볼 수 없다는 것에서 기인하며 차원축소가 이러한 것을 충족시키기 위해서라도 필요하다는 것을 보여준다.

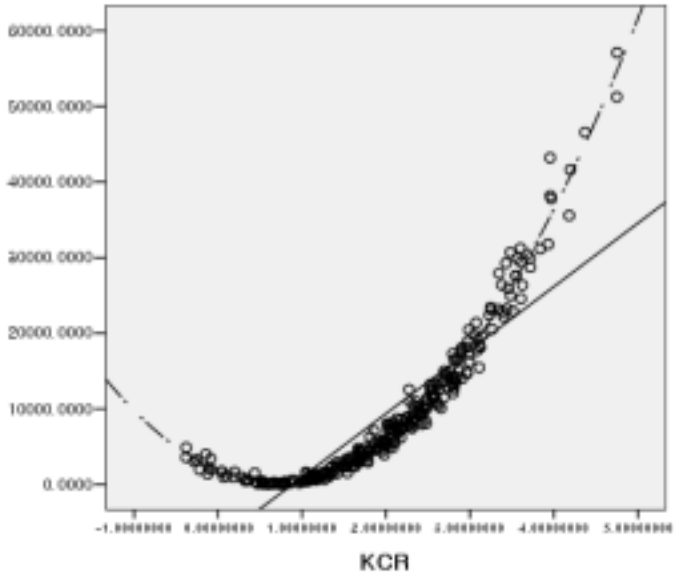
다음으로 커널 등고선회귀 방법에 의하여 차원을 축소하여 새로운 변수를 생성하였다. 앞에서 설명한 차원의 추정방법에 의하여 하나의 차원으로도 충분히 목표변수와의 관계 설명이 가능할 것으로 보이며, 새로이 생성된 변수(KCR)를 통하여 목표변수와의 관계를 산점도를 통하여 직접 확인할 수 있다(차원결정을 위한 결과는 부록을 참조).

[그림 2-1]에서 목표변수의 추정시 1차 형태보다는 2차 형태로 추정해야 더 적절하다는 것을 알 수 있다. 따라서 다음의 추정량으로 대체를 하였으며, KCR변수를 최근방대체를 위한 보조변수로 사용하였다.

$$\hat{y}=\hat{\beta_0}+\hat{\beta_1}KCR+\hat{\beta_2}KCR^2$$

<표 2-1>에서 커널 등고선회귀에 의해 차원축소를 한 경우의 대체가 기존 방법에 의한 대체보다 더 좋은 대체값을 추정할 수 있음을 알 수 있다. 이는 보조변수들의 정보의 손실 없이 차원을 축소함으로써 더 많

[그림 2-1] 모의실험 I 의 목표변수와 KCR변수의 산점도



〈표 2-1〉 모의실험 I 결과 (mean=8268.73)

대체방법 / 무응답률	통계량	M	R	KCRR	N	KCRN
10%	MAE_{mean}	134.88	74.88	22.40	173.72	31.75
	CV	1.93	1.13	0.34	2.61	0.48
20%	MAE_{mean}	205.66	118.08	35.32	263.19	51.36
	CV	3.26	1.80	0.55	3.89	0.78
30%	MAE_{mean}	268.01	148.08	50.45	339.33	72.72
	CV	4.25	2.24	0.76	5.05	1.09
40%	MAE_{mean}	339.92	181.56	66.16	395.26	92.96
	CV	5.27	2.81	1.03	6.07	1.39
50%	MAE_{mean}	394.83	229.45	84.58	481.98	120.94
	CV	6.31	3.47	1.28	7.23	1.76

은 정보의 사용과 더불어 직접적으로 목표변수와 보조변수와의 관계를 확인하여 모형을 추정할 수 있기 때문이다. 무응답률이 증가할수록 대체의 효율성은 조금씩 떨어지나, 차원축소를 통한 대체가 더 효율적인 것은 변함이 없다. 또한 목표변수와 매우 높은 상관도를 가지는 보조변수가 존재할 때에는 최근방대체 방법의 효율성은 더 좋아질 것이며, 커널 등고선회귀에 의해 차원축소를 한 경우의 최근방대체는 이보다 더 우수할 것으로 보인다. CV값을 통해 본 추정량의 안정성도 같은 패턴을 보여주고 있다.

2) $y=(X_1-X_3+X_4-X_5+X_6-X_9+X_{10})^3+\epsilon$ **모형 (모의실험 II)**

상관분석 및 변수선택 방법에 의한 결과 보조변수 X_3, X_4, X_5를 회귀대체를 위한 모형 추정에 사용하는 것이 가장 적절하다고 판단되며, 목표변수와 가장 상관도가 높은 X_4는 최근방대체를 위한 보조변수로 사용하였다(부록 참조). 결측된 값은 다음의 추정량으로 대체가 되었다.

$$\hat{y}=\hat{\beta_0}+\hat{\beta_1}X_3+\hat{\beta_2}X_4+\hat{\beta_3}X_5$$

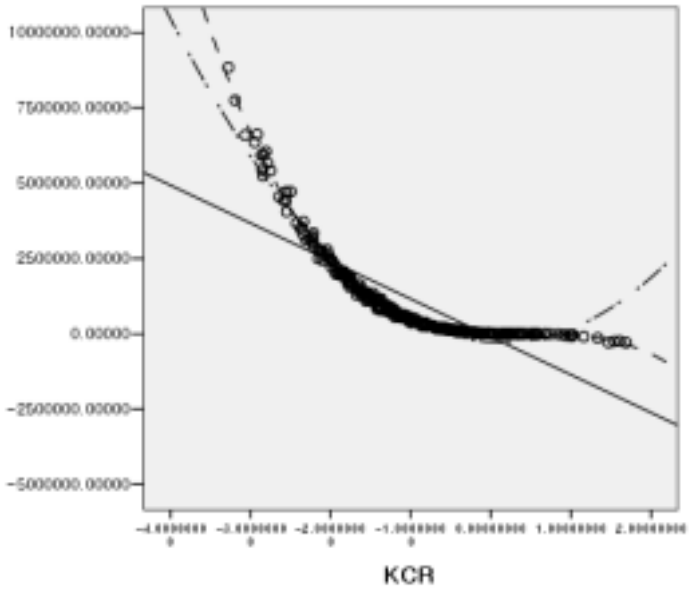
하나의 차원으로도 충분히 목표변수와의 관계 설명이 가능할 것으로 보이며(부록 참조), 새로이 생성된 변수(KCR)를 통하여 목표변수와의 관계를 산점도를 통하여 직접 확인하였다.

[그림 2-2]에서 목표변수의 추정시 1차나 2차의 형태보다는 3차의 형태로 추정하는 것이 더 적절하다는 것을 알 수 있다. 따라서 다음의 추정량으로 대체를 하였으며, KCR변수를 최근방대체를 위한 보조변수로 사용하였다.

$$\hat{y}=\hat{\beta_0}+\hat{\beta_1}KCR+\hat{\beta_2}KCR^2+\hat{\beta_3}KCR^3$$

<표 2-2>에서도 커널 등고선회귀에 의해 차원축소를 한 경우의 대체가 기존방법에 의한 대체보다 더 좋은 대체값을 추정할 수 있음을 알 수 있다. 이는 앞의 모의실험에서와 같은 이유로 볼 수 있다.

[그림 2-2] 모의실험 II의 목표변수와 KCR변수의 산점도



〈표 2-2〉 모의실험 II 결과 (mean=1112683.32)

대체방법 / 무응답률	통계량	M	R	KCRR	N	KCRN
10%	MAE_{mean}	27384.55	16272.79	2530.33	31732.43	6346.88
	CV	3.17	1.87	0.28	3.66	0.91
20%	MAE_{mean}	39411.63	25552.23	4179.74	51449.01	10652.18
	CV	4.55	2.84	0.48	5.75	1.25
30%	MAE_{mean}	49128.27	32439.85	5975.31	63341.52	14978.57
	CV	5.59	3.63	0.68	7.11	1.57
40%	MAE_{mean}	61755.47	41713.76	7582.62	76426.43	16843.94
	CV	7.01	5.84	0.87	9.19	2.03
50%	MAE_{mean}	87435.93	58931.89	9762.85	87588.77	20463.50
	CV	9.69	7.43	1.04	11.06	2.74

3) $y=(X_1-X_3+X_4-X_5+X_6-X_9+X_{10})^1+\epsilon$ 모형 (모의실험 III)

상관분석 및 변수선택 방법에 의한 결과 보조변수 X_3, X_4, X_5, X_{10}을 회귀대체를 위한 모형 추정에 사용하는 것이 가장 적절하다고 판단되며, 목표변수와 가장 상관도가 높은 X_4는 최근방대체를 위한 보조변수로 사용하였다(부록 참조). 결측된 값은 다음의 추정량으로 대체가 되었다.

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1X_3+\hat{\beta}_2X_4+\hat{\beta}_3X_5+\hat{\beta}_4X_{10}$$

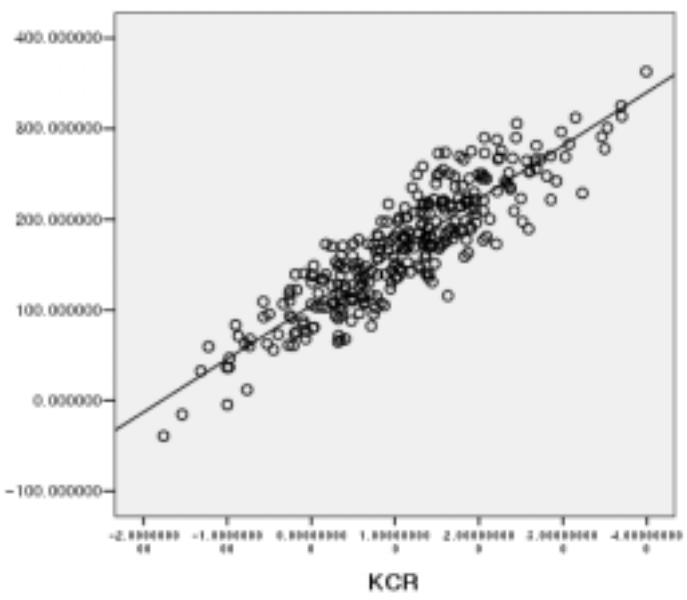
하나의 차원으로도 충분히 목표변수와의 관계 설명이 가능할 것으로 보이며, 새로이 생성된 변수(KCR)를 통하여 목표변수와의 관계를 산점도를 통하여 직접 확인하였다.

[그림 2-3]에서 목표변수를 추정할 때 1차 선형 형태로 추정하는 것이 가장 적절하다는 것을 알 수 있다. 따라서 다음의 추정량으로 대체를 하였으며, KCR변수를 최근방대체를 위한 보조변수로 사용하였다.

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1KCR$$

<표 2-3>에서도 커널 등고선회귀에 의해 차원축소를 한 경우의 대체가 기존 방법에 의한 대체보다 더 좋은 효율을 준다. 그러나 이 경우에는 큰 차이를 보이지 않는 것을 알 수 있다. 이는 실제 모형이 선형이기 때문에 직접적으로 목표변수와 보조변수의 관계를 확인하지 않고도 기존의 방법으로 좋은 대체가 가능한 것이다. 하지만 현실적으로 이러한 선형모형보다는 더 복잡한 모형들이 상당히 많이 존재한다. 따라서 다양한 모형들에 대해서 차원축소를 통하여 목표변수와 보조변수의 관계를 확인하여 적절한 대체모형을 찾을 수 있는 커널 등고선회귀를 이용한 방법이 더 효율적인 대체를 하기 위한 하나의 방법이라고 할 수 있을 것이다.

[그림 2-3] 모의실험 III의 목표변수와 KCR변수의 산점도



〈표 2-3〉 모의실험 III 결과 (mean=174.31)

대체방법 / 무응답률	통계량	M	R	KCRR	N	KCRN
10%	MAE_{mean}	0.9129	0.5279	0.4666	1.2382	0.6246
	CV	0.65	0.38	0.32	0.85	0.44
20%	MAE_{mean}	1.4294	0.7733	0.6525	1.7853	0.9449
	CV	0.98	0.55	0.47	1.27	0.67
30%	MAE_{mean}	1.8055	1.0468	0.9077	2.3086	1.1981
	CV	1.26	0.74	0.65	1.66	0.86
40%	MAE_{mean}	2.4406	1.3137	1.1416	2.7390	1.4753
	CV	1.71	0.94	0.81	1.97	1.05
50%	MAE_{mean}	3.1256	1.5846	1.3961	3.2345	1.8192
	CV	2.22	1.13	1.06	2.31	1.32

나. 감마분포로 보조변수 생성

1) $y=(X_1-X_3+X_4-X_5+X_6-X_9+X_{10})^2+\epsilon$ 모형 (모의실험 Ⅳ)

앞에서 설명한 바와 같이 보조변수의 왜도(skewness)를 고려한 모의실험으로 감마분포에 의하여 보조변수를 생성하였다. 또한 무응답 bias를 고려해 목표변수의 중위수 이상의 값에 대해서 무응답을 발생시켰다. 상관분석 및 변수선택 방법에 의한 결과 보조변수 X_1, X_4, X_6, X_{10}을 회귀대체를 위한 모형 추정에 사용하는 것이 가장 적절하다고 판단되며, 목표변수와 가장 상관도가 높은 X_4는 최근방대체를 위한 보조변수로 사용하였다(부록 참조). 결측된 값은 다음의 추정량으로 대체가 되었다.

$$\hat{y}=\hat{\beta_0}+\hat{\beta_1}X_1+\hat{\beta_2}X_4+\hat{\beta_3}X_6+\hat{\beta_4}X_{10}$$

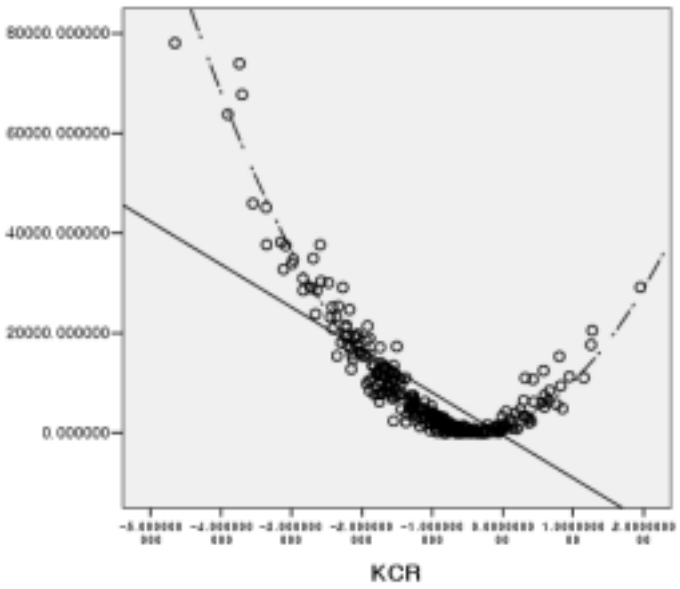
하나의 차원으로도 충분히 목표변수와의 관계 설명이 가능할 것으로 보이며, 새로이 생성된 변수(KCR)를 통하여 목표변수와의 관계를 산점도를 통하여 확인하였다.

[그림 2-4]에서 알 수 있듯이 목표변수를 2차 형태로 추정해야 한다. 따라서 다음의 추정량으로 대체를 하였으며, KCR변수를 최근방대체를 위한 보조변수로 사용하였다.

$$\hat{y}=\hat{\beta_0}+\hat{\beta_1}KCR+\hat{\beta_2}KCR^2$$

<표 2-4>에서도 커널 등고선회귀에 의해 차원축소를 한 경우의 대체가 기존 방법에 의한 대체보다 더 좋은 효율을 준다. 또한 이 경우에는 목표변수의 큰 값에 대해서 무응답이 발생되었으므로 무응답 대체방법에 따라 평균 추정에 미치는 영향이 더 커지게 된다. 커널 등고선회귀에 의해 차원축소를 한 경우의 대체는 이러한 측면에서 매우 안정적인 방법이라고 할 수 있을 것이다. 또한 기존의 방법에 비해 보조변수의 분포나 무응답 bias에도 큰 영향을 받지 않는 것을 알 수 있다.

[그림 2-4] 모의실험 IV의 목표변수와 KCR변수의 산점도



〈표 2-4〉 모의실험 IV 결과 (mean=9211.5)

대체방법 / 무응답률	통계량	M	R	KCRR	N	KCRN
3%	MAE_{mean}	254.07	162.80	49.78	263.84	56.88
	CV	1.66	1.41	0.44	2.05	0.72
5%	MAE_{mean}	423.70	267.62	58.12	428.98	6,5.90
	CV	2.28	1.99	0.58	2.77	0.85
7%	MAE_{mean}	581.83	367.28	78.89	559.23	98.09
	CV	2.49	2.14	0.76	3.10	1.05
10%	MAE_{mean}	889.71	562.80	107.76	861.11	132.18
	CV	3.52	2.88	0.93	3.87	1.32
15%	MAE_{mean}	1,384.61	875.27	177.18	1,316.24	201.99
	CV	4.42	3.59	1.24	5.28	1.62

2) $y=(X_1-X_3+X_4-X_5+X_6-X_9+X_{10})^3+\epsilon$ 모형 (모의실험 V)

상관분석 및 변수선택 방법에 의한 결과 보조변수 X_1, X_4, X_6, X_{10}을 회귀대체를 위한 모형 추정에 사용하는 것이 가장 적절하다고 판단되며, 목표변수와 가장 상관도가 높은 X_4는 최근방대체를 위한 보조변수로 사용하였다(부록 참조). 결측된 값은 다음의 추정량으로 대체가 되었다.

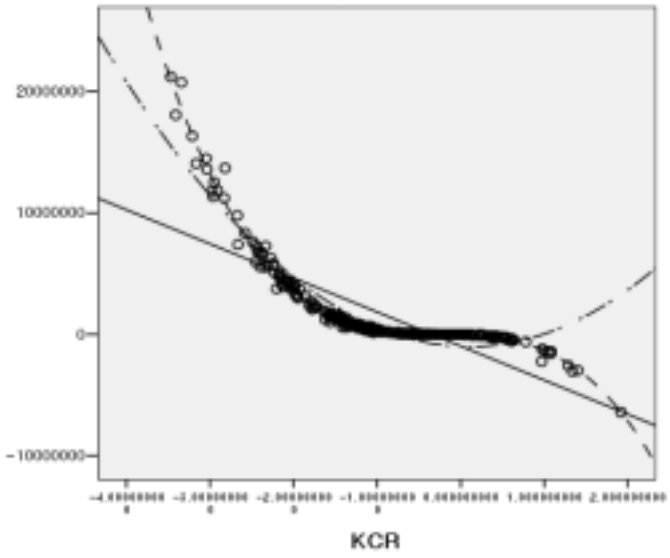
$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_4+\hat{\beta}_3X_6+\hat{\beta}_4X_{10}$$

모의실험 V 역시 하나의 차원으로도 충분히 목표변수와의 관계 설명이 가능할 것으로 보이며(부록 참조), 새로이 생성된 변수(KCR)를 통하여 목표변수와의 관계를 산점도를 통하여 확인하였다.

[그림 2-5]에서 목표변수의 추정시 1차나 2차의 형태보다는 3차 형태로 추정하는 것이 더 적절하다는 것을 알 수 있다. 따라서 다음의 추정량으로 대체를 하였으며, KCR변수를 최근방대체를 위한 보조변수로 사용하였다.

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1KCR+\hat{\beta}_2KCR^2+\hat{\beta}_3KCR^3$$

[그림 2-5] 모의실험 V의 목표변수와 KCR변수의 산점도



<표 2-5>의 결과도 정규분포로 보조변수를 생성한 경우와 유사하다는 것을 알 수 있다. 이는 앞의 모의실험에서와 같은 이유이다.

〈표 2-5〉 모의실험 V 결과 (mean=1264829.86)

무응답률 \ 대체방법	통계량	M	R	KCRR	N	KCRN
3%	MAE_{mean}	47,198.5	31,295.5	5,386.2	45,288.8	8,339.1
	CV	3.26	3.02	0.46	3.89	0.72
5%	MAE_{mean}	78,282.3	49,518.7	9,553.9	69,773.2	13,471.9
	CV	4.20	3.95	0.81	4.81	1.03
7%	MAE_{mean}	108,146.4	58,200.6	13,259.8	88,874.5	19,653.4
	CV	5.07	4.64	1.12	5.76	1.65
10%	MAE_{mean}	172,026.5	94,261.9	17,461.2	138,201.1	24,603.2
	CV	6.77	6.19	1.47	6.97	2.11
15%	MAE_{mean}	258,704.8	129,413.6	23,219.5	196,886.4	32,279.9
	CV	8.61	8.01	1.97	9.24	2.72

3) $y=(X_1-X_3+X_4-X_5+X_6-X_9+X_{10})^1+\epsilon$ 모형 (모의실험 VI)

상관분석 및 변수선택 방법에 의한 결과 보조변수 X_1, X_3, X_4, X_5, X_{10}을 회귀대체를 위한 모형 추정에 사용하는 것이 가장 적절하다고 판단되며, 목표변수와 가장 상관도가 높은 X_4는 최근방대체를 위한 보조변수로 사용하였다(부록 참조). 결측된 값은 다음의 추정량으로 대체가 되었다.

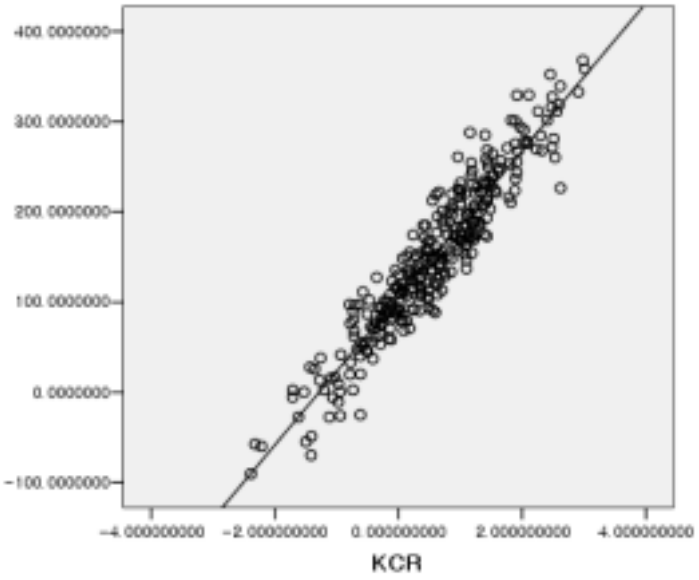
$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_3+\hat{\beta}_3X_4+\hat{\beta}_4X_5+\hat{\beta}_5X_{10}$$

하나의 차원으로도 충분히 목표변수와의 관계 설명이 가능할 것으로 보이며(부록 참조), 새로이 생성된 변수(KCR)를 통하여 목표변수와의 관계를 산점도를 통하여 직접 확인하였다.

[그림 2-6]에서 목표변수의 추정시 1차 선형 형태로 추정하는 것이 가장 적절하다는 것을 알 수 있다. 따라서 다음의 추정량으로 대체를 하였으며, KCR변수를 최근방대체를 위한 보조변수로 사용하였다.

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1KCR$$

[그림 2-6] 모의실험 VI의 목표변수와 KCR변수의 산점도



〈표 2-6〉 모의실험 VI 결과 (mean=170.90)

대체방법 / 무응답률	통계량	M	R	KCRR	N	KCRN
3%	MAE_{mean}	1.5588	0.4398	0.4127	1.0542	0.4723
	CV	0.22	0.18	0.17	1.39	0.25
5%	MAE_{mean}	2.6017	0.6942	0.6452	1.8792	0.6824
	CV	0.28	0.25	0.24	0.49	0.33
7%	MAE_{mean}	3.7624	0.9783	0.9074	2.6776	0.9312
	CV	0.36	0.28	0.27	0.62	0.30
10%	MAE_{mean}	5.5067	1.4443	1.3083	3.9694	1.3431
	CV	0.53	0.35	0.32	0.80	0.33
15%	MAE_{mean}	8.8570	2.2811	1.9261	6.5147	2.0047
	CV	0.74	0.45	0.41	0.99	0.42

<표 2-6>의 결과 역시 모의실험 Ⅲ과 같은 결론을 내릴 수 있을 것이다. 앞의 모의실험들의 결과를 종합해 보면 일반적으로 커널 등고선회귀에 의해 차원축소를 한 경우의 대체가 기존 방법에 의한 대체보다 더 좋은 효율을 준다는 것을 보였다. 이는 많은 보조변수의 정보를 활용할 뿐만 아니라 목표변수의 관계를 저차원에서 확인이 가능할 경우 대체모형을 더 정확하게 추정할 수 있게 되는 이점을 최대한 이용한 결과라 하겠다. 마지막으로 조금 더 복잡한 모형에 대하여 적용해 보고 본 모의실험을 마치고자 한다.

다. 복잡한 모형에서의 검토(표준정규분포로 보조변수 생성)

앞의 모의실험에서보다 조금 더 복잡한 모형을 고려하여 실험을 하고자 한다. 모든 과정은 동일하며 모형만 다음과 같이 설정하였다. 이 모형은 2차와 3차 형태를 동시에 가지고 있는 것을 알 수 있다.

$$y = (100*(X_1 + X_4 + X_5 + X_9 + X_{10}))^2 + (20*(X_2 + X_7))^3 + \epsilon$$ **모형 (모의실험 Ⅶ)**

상관분석 및 변수선택 방법에 의한 결과 보조변수 X_2, X_7이 회귀대체를 위한 모형 추정에 사용하는 것이 적절하다고 판단되며, 목표변수와 가장 상관도가 높은 X_7은 최근방대체를 위한 보조변수로 사용하였다(부록 참조). 결측된 값은 다음의 추정량으로 대체가 되었다.

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} X_2 + \hat{\beta_2} X_7$$

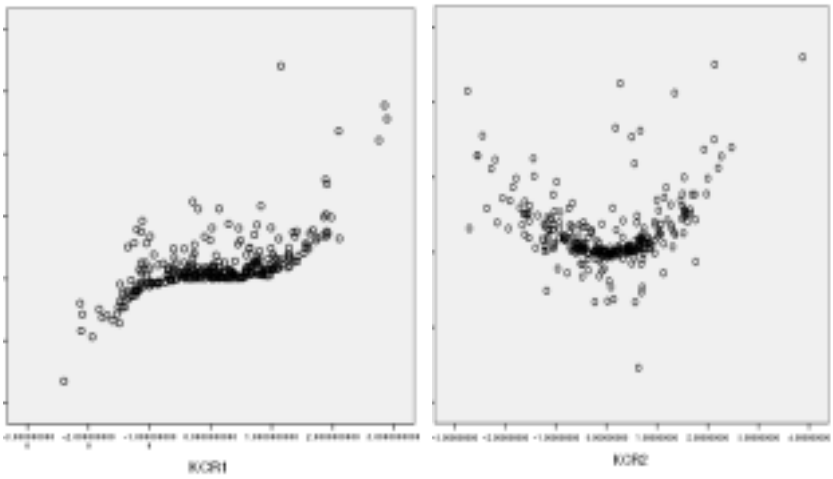
앞의 모의실험과는 달리 차원선택의 결과 2차원 구조로 설명이 가능할 것으로 보인다(부록 참조). 이는 목표변수가 2차와 3차 형태를 동시에 가지고 있기 때문에 두 방향에서 모형을 구축하는 것으로 정확하게 차원을 결정해 주고 있다는 것을 알 수 있다. 새로이 생성된 두 변수(KCR1, KCR2)를 통하여 목표변수와의 관계를 산점도를 통하여 직접 확인하였다.

[그림 2-7]에서 알 수 있듯이 KCR1 변수는 목표변수와 3차 곡선형태의 관계를 보이며, KCR2 변수는 2차 곡선형태가 있는 것으로 보인다.

따라서 다음의 추정량으로 대체를 하였으며, 목표변수와 가장 높은 상관관계가 있는 KCR1 변수를 최근방대체를 위한 보조변수로 사용하였다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 KCR_1^3 + \hat{\beta}_2 KCR_2^2$$

[그림 2-7] 모의실험 Ⅶ의 목표변수와 KCR1과 KCR2 변수의 산점도



〈표 2-7〉 모의실험 Ⅶ 결과 (mean=44599.31)

대체방법 / 무응답률	통계량	M	R	KCRR	N	KCRN
10%	MAE_{mean}	1,682.14	1,137.58	398.61	1,783.03	1,507.33
	CV	4.62	3.11	1.23	4.95	4.28
20%	MAE_{mean}	2,656.30	1,724.94	748.92	2,628.14	2,201.80
	CV	7.15	4.83	2.39	7.26	6.25
30%	MAE_{mean}	3,360.11	2,262.17	1,075.42	3,383.20	3,015.37
	CV	9.26	6.83	3.35	9.43	8.36
40%	MAE_{mean}	3,880.28	2,830.61	1,642.46	4,183.56	3,625.02
	CV	10.96	7.94	5.96	11.39	9.27
50%	MAE_{mean}	5,155.04	3,551.32	2,233.84	5,234.65	4,450.71
	CV	14.02	9.26	6.60	14.98	12.67

<표 2-7>의 결과도 앞의 모의실험들의 경우와 유사하다는 것을 알 수 있다. 실제모형이 선형에서 멀어질수록 기존의 대체방법은 많은 보조변수를 사용할 때 정확한 모형을 추정하기가 힘들다. 그러나 커널 등고선회귀 방법으로 차원을 축소하여 이를 통해 생성된 보조변수를 이용하면 많은 정보의 사용과 저차원으로 인한 모형구축의 용이성이 함께 보장되어 기존의 대체방법에 비해서 많은 경우 더 좋은 대체를 할 수 있다는 것을 본 모의실험을 통해 알 수 있다.

제5절 결론 및 향후 연구의 방향

본 연구에서는 항목 무응답 대체를 위한 새로운 방법으로 차원축소 기법인 커널 등고선회귀를 이용한 방법을 살펴보았다. 항목 무응답을 대체하는 경우 소득이나 매출액과 같은 수치형 자료일 때에는 주로 회귀대체, 최근방대체, 과거 자료값의 대체, 과거 자료값과의 증감을 고려한 대체 등을 일반적으로 적용하고 있다. 이러한 방법들은 많은 부분에서 적절한 대체방법일 것이다. 그러나 보조정보를 사용한다는 측면에서는 한계가 있을 것이다. 많은 보조정보의 사용은 추정의 복잡성을 가져오며 적은 보조변수의 사용은 추정의 효율성이 저하될 수 있기 때문이다. 본 연구에서는 차원축소에 의한 새로운 보조변수의 생성으로 이러한 문제가 극복될 수 있음을 보여준다. 즉, 많은 보조변수를 사용하더라도 차원의 수는 크지 않게 되므로 추정의 복잡성과 효율성 문제가 동시에 해결 가능하다는 것이다. 앞의 모의실험의 결과를 통하여 차원축소에 대한 이점과 대체방법에의 적용가능성이 확인되었으리라 판단된다. 따라서 본 연구의 결과를 토대로 향후 연구되어야 할 부분을 제시하고자 한다.

첫째, 제시된 새로운 방법을 이용하여 실제자료에 적용함으로써 실제조사의 적용타당성 검토가 이루어져야 할 것이다. 이 방법은 주로 수치형변수가 많은 조사에 주로 이용가능하다. 농가경제조사나 어가경제조사 등이 적합할 것으로 본다. 둘째, 무응답 대체군 형성에 관한 연구에도 응용할 수 있을 것이다. 무응답률이 증가하면 대체군 형성의 필요

성은 증가할 것이며 좋은 대체군을 이용하면 더욱 효율적인 대체가 가능하다. 대체군 형성과정에서 대체군 내의 보조변수의 수도 중요한 문제가 될 것이다. 이 과정에서 수치형자료의 보조변수들을 축소시키는데 본 연구의 결과를 이용할 수도 있을 것으로 본다.

대부분의 통계조사에서 무응답은 현실적으로 존재한다. 무응답률을 최대한 낮추는 것이 바람직하지만 시간·공간적인 여건과 비용의 제약으로 무응답의 효과를 최대한 보정하는 것이 조사의 통계품질을 향상시키는 방법이 될 것이다. 무응답 대체방법은 통계조사에 따라 적절히 선택되어야 하며, 각 조사에 맞는 새로운 대체방법도 계속적으로 연구가 되어야 할 것이다. 이러한 연구를 통하여 무응답의 효과를 완전히 없앨 수는 없지만 사후적인 처리를 통하여 무응답 효과를 최소화할 수 있을 것으로 본다. 본 연구에서 제시된 방법 역시 무응답 대체를 위한 하나의 새로운 방법으로 적절한 조사에 잘 적용되어 통계품질을 향상시키는 데 일조할 수 있기를 기대한다.

참고문헌

김규성(2000), “무응답 대체 방법과 대체 효과”, 「조사연구」, 제1권 2호, pp.1-14.

김규성(2000), “표본 대체 방법과 대체자료의 합리적 이용”, 한국은행 지원논문.

김규성・이기재・김진(2005), “농어가경제조사에서 가중핫덱 무응답 대체방법의 활용”, 「응용통계연구」, 제18권 2호, pp.311-328.

김영원・조선경(1996), “표본조사에서 항목 무응답 대체 방법”, 「한국통계학회논문집」, 제3권 3호, pp.145-159.

김재광・한근식・윤연옥(2004), “가계조사 무응답 처리기법 연구”, 통계청, 「통계연구」, 제9권 제1호, pp.79-102.

김진(2004), “농가경제조사에 대한 대체법 비교”, 통계청, 「통계연구」, 제9권 제2호, pp.133-145.

이진희・김진・이기재(2006), “표본조사에서 공간 변수를 이용한 결측대체의 효율성 비교”, 「응용통계연구」, 제19권 1호, pp.57-67.

이학배・최필근(2004), “회귀그래픽 방법에서의 다중공선성”, *Journal of the Korean Data Analysis Society*, Vol. 6, No. 3, pp.849-860.

조사통계연구회(2000), 「무응답 오차」, 자유아카데미.

최통진(2006), “농림어업총조사를 위한 무응답 보정에 관한 연구”, 석사학위논문.

통계교육원(2005), “무응답처리 실무론”.

Altman, N. S.(1992), “An Introduction to Kernel and Nearest Neighbor Nonparametric Regression”, *The American Statistician*, 46, pp.175-185.

Choi, P. K.(2006), “Kernel Contour Regression for Dimension Reduction”, Doctoral Dissertation.

Cook, R. D.(1998), *Regression Graphics*, Wiley, New York.

Cook, R. D. and X. Yin(2001), “Dimension Reduction and Visualization in Discriminant Analysis”, *Australian and New Zealand Journal of Statistics*, 43, pp.147-199.

Li, B., H. Zha, and F. Chiaromonte(2005), “Contour Regression: A General Approach to Dimension Reduction”, *The Annals of Statistics*, Vol. 33, No. 4, pp.1580-1616.

Rubin, D. B. and J. A. Little(1986), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

< 부 록 >

1. 모의실험 I 을 위한 보조변수선택 및 차원결정 결과

▸ stepwise 방법에 의한 보조변수 선택결과

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.85E+009	1	7849418497.74	111.122	.000
	Residual	2.11E+010	298	70637838.503		
	Total	2.89E+010	299			
2	Regression	1.56E+010	2	7815910412.72	174.961	.000
	Residual	1.33E+010	297	44672301.503		
	Total	2.89E+010	299			
3	Regression	1.90E+010	3	6336768628.04	189.670	.000
	Residual	9.89E+009	296	33409420.566		
	Total	2.89E+010	299			
4	Regression	2.09E+010	4	5214869594.41	191.341	.000
	Residual	8.04E+009	295	27254291.505		
	Total	2.89E+010	299			
5	Regression	2.23E+010	5	4464451367.94	199.559	.000
	Residual	6.58E+009	294	22371556.231		
	Total	2.89E+010	299			
6	Regression	2.29E+010	6	3811791356.60	185.255	.000
	Residual	6.03E+009	293	20575925.707		
	Total	2.89E+010	299			
7	Regression	2.30E+010	7	3281991730.91	161.730	.000
	Residual	5.93E+009	292	20292987.176		
	Total	2.89E+010	299			

* 모형5에 해당되는 변수: X_1, X_3, X_4, X_5, X_{10}.

▸ 커널 등고선회귀의 차원 결정

고유값 1	1.7015	고유값 6	0.0130
2	0.4037	7	-0.0247
3	0.2785	8	-0.1123
4	0.1774	9	-0.2032
5	0.0860	10	-0.3287

2. 모의실험 II를 위한 보조변수선택 및 차원결정 결과

▸ stepwise 방법에 의한 보조변수 선택결과

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.85E+014	1	1.847E+014	107.416	.000
	Residual	5.12E+014	298	1.719E+012		
	Total	6.97E+014	299			
2	Regression	3.06E+014	2	1.529E+014	116.050	.000
	Residual	3.91E+014	297	1.317E+012		
	Total	6.97E+014	299			
3	Regression	3.88E+014	3	1.292E+014	123.548	.000
	Residual	3.10E+014	296	1.046+012		
	Total	6.97E+014	299			
4	Regression	4.31E+014	4	1.078E+014	119.675	.000
	Residual	2.66E+014	295	900972926871		
	Total	6.97E+014	299			
5	Regression	4.61E+014	5	9.218E+013	114.749	.000
	Residual	2.36E+014	294	803326355999		
	Total	6.97E+014	299			
6	Regression	4.82E+014	6	8.027E+013	109.175	.000
	Residual	2.15E+014	293	735280102913		
	Total	6.97E+014	299			

* 모형3에 해당되는 변수: X_3, X_4, X_5.

▸ 커널 등고선회귀의 차원 결정

고유값 1	1.8423	고유값 6	0.0086
2	0.4637	7	-0.0921
3	0.3780	8	-0.1342
4	0.2493	9	-0.2668
5	0.1245	10	-0.3848

3. 모의실험 III을 위한 보조변수선택 및 차원결정 결과

▸ stepwise방법에 의한 보조변수 선택결과

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	380887.283	1	380887.283	120.370	.000
	Residual	942964.878	298	3164.312		
	Total	1323852.2	299			
2	Regression	7743826.141	2	371913.071	190.437	.000
	Residual	580026.020	297	1952.950		
	Total	1323852.2	299			
3	Regression	894517.500	3	298172.500	205.572	.000
	Residual	429334.661	296	1450.455		
	Total	1323852.2	299			
4	Regression	993484.874	4	248371.219	221.782	.000
	Residual	330367.287	295	1119.889		
	Total	1323852.2	299			
5	Regression	1024342.6	5	204868.511	201.100	.000
	Residual	299509.609	294	1018.740		
	Total	1323852.2	299			
6	Regression	1045253.8	6	174208.973	183.214	.000
	Residual	278598.322	293	950.848		
	Total	1323852.2	299			
7	Regression	1049530.9	7	149932.992	159.596	.000
	Residual	274321.219	292	939.456		
	Total	1323852.2	299			

* 모형4에 해당되는 변수 : X_3, X_4, X_5, X_{10}.

▸ 커널 등고선회귀의 차원 결정

고유값 1	1.6762	고유값 6	-0.0218
2	0.2729	7	-0.0843
3	0.1794	8	-0.1487
4	0.1094	9	-0.2146
5	0.0374	10	-0.2888

4. 모의실험 Ⅳ를 위한 보조변수선택 및 차원결정 결과

▶ stepwise방법에 의한 보조변수 선택결과

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.80E+009	1	9799060456.88	59.094	.000
	Residual	4.94E+010	298	165821131.797		
	Total	5.92E+010	299			
2	Regression	1.74E+010	2	8688789501.90	61.683	.000
	Residual	4.18E+010	297	140862554.642		
	Total	5.92E+010	299			
3	Regression	2.34E+010	3	7798804099.49	64.451	.000
	Residual	3.58E+010	296	121004545.385		
	Total	5.92E+010	299			
4	Regression	2.69E+010	4	6717921051.38	61.276	.000
	Residual	3.23E+010	295	109634147.549		
	Total	5.92E+010	299			
5	Regression	2.89E+010	5	5785483202.80	56.162	.000
	Residual	3.03E+010	294	103014767.749		
	Total	5.92E+010	299			
6	Regression	3.08E+010	6	5132726824.22	52.921	.000
	Residual	2.84E+010	293	96987702.345		
	Total	5.92E+010	299			
7	Regression	3.15E+010	7	4495115550.56	47.303	.000
	Residual	2.77E+010	292	95027222.186		
	Total	5.92E+010	299			

* 모형4에 해당되는 변수 : X_1, X_4, X_6, X_{10}.

▶ 커널 등고선회귀의 차원 결정

고유값 1	1.6076	고유값 6	0.0886
2	0.6627	7	-0.0627
3	0.5862	8	-0.1122
4	0.4205	9	-0.2386
5	0.2461	10	-0.3918

5. 모의실험 Ⅴ를 위한 보조변수선택 및 차원결정 결과

▸ stepwise방법에 의한 보조변수 선택결과

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.07E+014	1	7.070E+014	52.488	.000
	Residual	4.01E+015	298	1.347E+013		
	Total	4.72E+015	299			
2	Regression	1.30E+015	2	6.518E+014	56.654	.000
	Residual	3.42E+015	297	1.151E+013		
	Total	4.72E+015	299			
3	Regression	1.83E+015	3	6.086E+014	62.234	.000
	Residual	2.89E+015	296	7.780E+012		
	Total	4.72E+015	299			
4	Regression	2.21E+015	4	5.525E+014	64.911	.000
	Residual	2.51E+015	295	8.511E+012		
	Total	4.72E+015	299			
5	Regression	2.47E+015	5	4.943E+014	64.598	.000
	Residual	2.25E+015	294	7.651E+012		
	Total	4.72E+015	299			
6	Regression	2.57E+015	6	4.276E+014	58.137	.000
	Residual	2.16E+015	293	7.355E+012		
	Total	4.72E+015	299			
7	Regression	2.64E+015	7	3.771E+014	52.895	.000
	Residual	2.08E+015	292	7.128E+012		
	Total	4.72E+015	299			

* 모형4에 해당되는 변수 : X_1, X_4, X_6, X_{10}.

▸ 커널 등고선회귀의 차원 결정

고유값 1	1.8909	고유값 6	0.1407
2	0.6784	7	-0.0396
3	0.5809	8	-0.1755
4	0.4311	9	-0.3486
5	0.3128	10	-0.4521

6. 모의실험 VI을 위한 보조변수선택 및 차원결정 결과

▶ stepwise방법에 의한 보조변수 선택결과

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	328859.396	1	328859.396	108.641	.000
	Residual	902055.996	298	3027.034		
	Total	1230915.4	299			
2	Regression	604290.129	2	302145.065	143.207	.000
	Residual	626625.263	297	2109.849		
	Total	1230915.4	299			
3	Regression	738924.449	3	246308.150	148.188	.000
	Residual	491990.943	296	1662.132		
	Total	1230915.4	299			
4	Regression	846191.100	4	211547.775	162.211	.000
	Residual	384724.292	295	1304.150		
	Total	1230915.4	299			
5	Regression	924085.247	5	184817.049	177.089	.000
	Residual	306830.145	294	1043.640		
	Total	1230915.4	299			
6	Regression	924085.247	6	159915.681	172.629	.000
	Residual	306830.145	293	926.353		
	Total	1230915.4	299			

* 모형7에 해당되는 변수 : X_1, X_3, X_4, X_5, X_{10}.

▶ 커널 등고선회귀의 차원 결정

고유값 1	1.7745	고유값 6	0.0311
2	0.4744	7	-0.0557
3	0.3303	8	-0.1443
4	0.2224	9	-0.2366
5	0.1263	10	-0.3570

7. 모의실험 Ⅶ을 위한 보조변수선택 및 차원결정 결과

▸ stepwise방법에 의한 보조변수 선택결과

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.83E+011	1	582853994513	80.254	.000
	Residual	2.16E+012	298	7262610770.64		
	Total	2.75E+012	299			
2	Regression	1.17E+012	2	584775354924	110.093	.000
	Residual	1.58E+012	297	5311654189.62		
	Total	2.75E+012	299			

* 모형2에 해당되는 변수 : X_2, X_7.

▸ 커널 등고선회귀의 차원 결정

고유값 1	1.4275	고유값 6	0.0433
2	1.2119	7	-0.0689
3	0.4359	8	-0.1950
4	0.2806	9	-0.2882
5	0.1586	10	-0.4118