碩士學位論文

로지스틱 회귀분석에서 희소주성분회귀법의 효율성 연구

韓國外國語大學校 大學院
統 計 學 科
朴 允 美

碩士學位論文

로지스틱 회귀분석에서 희소주성분회귀법의 효율성 연구

指導 李 碩 浩 敎授

이 論文을 碩士學位請求論文으로 提出합니다.

2014年 月

韓國外國語大學校 大學院
統 計 學 科
朴 允 美

이 論文을 朴允美의 碩士學位論文으로 認定함

2014年 月 日

審 査 委 員 _____ (인)

審 査 委 員 _____ (인)

審 査 委 員 _____ (인)

韓國外國語大學校 大學院

요약

본 연구는 주성분 로지스틱 회귀분석에서 이용되는 주성분을 Lasso등 여러 희소벌점함수를 이용한 변수선택을 통해 추정량의 크기를 축소시키는 동시에 반응변수와 연관성이 높은 몇 개의 주성분만을 선택하여 추정의 정확도를 높이는 방안을 제시한다. 다양한 상황에서의 회귀분석 문제에서 기존의 방법론과의 비교를 컴퓨터 모의실험을 통해 수행하였으며, 일반 로지스틱 회귀분석보다 우수한 성능을 보임을 확인하였다.

목 차

표 목 차

그 림 목 차

1 서 론

1.1 연구배경 및 목적

유전자 정보(genetic information)를 활용한 생물통계 분야에서의 질병발현 원인연구나 금융 및 신용거래 정보(credit data)를 이용하여 개인의 신용도를 점수화하는 신용평점모형 개발 등 다양한 분야에서 로지스틱 회귀분석법(logistic regression)이 빈번하게 활용된다. 유전자 정보나 신용거래 자료처럼 반응변수(response variable)가 이항자료(binary data)로 되어있을 때 로지스틱 분석법이 주로 고려되며, 여러 설명변수와 반응변수간의 관계를 모형화 하는 로지스틱 회귀분석이나 추정된 모형에 의해 자료를 두 개의 범주로 분류하는 로지스틱 분류법을 적용한 다양한 연구들이 활발하게 진행 중이다.

하지만 최근 급속한 기술적 발전 및 저장 용량의 비용감소로 인해 대용량 데이터(massive data, Big data)의 규모가 단순히 관측치 수의 증가뿐만 아니라 설명변수의 수의 증가로 확대되었다. 즉, 자료의 개수보다 설명변수의 수가 더 많은 경우도 빈번하게 발생하고, 그 설명변수들간의 강한 상관관계가 나타나는 것이 일반적이다. 이처럼 설명변수간에 높은 상관성을 가지는 경우를 다중공선성(multicollinearity)이라고 한다. 다중공선성이 존재하는 경우 로지스틱 회귀모형을 그대로 적용시키게 되면, 왜곡된 회귀계수의 추정이나 잘못된 분류의 결과를 가져오게 된다. 또한, 고차원 자료의 경우 고려해야 하는 설명변수의 수가 많아지므로 로지스틱 회귀모형

적합 시 해석상(interpretability)의 문제가 발생한다. 다시 말해, 고려하는 설명변수에 대한 적절한 선택과정이 없다면 상호간의 연관성이 높은 설명변수들이 포함되게 되고, 해석이 어려운 문제들이 발생하게 된다. 따라서, 본 논문에서는 위와 같은 문제점을 해결하기 위한 추정기법에 관해 연구하고자 한다.

1.2 연구방법 및 구성

일반적으로 로지스틱 회귀모형에서의 모수 추정은 관측된 자료가 발생할 확률을 최대화 하는 최대우도추정법(maximum likelihood estimation)을 이용한다. 그러나 다중공선성이 존재하는 경우 최대우도추정량의 분산이 커지게 되어 추정값을 신뢰할 수 없다. 따라서 차원축소법(dimension reduction) 및 축소추정법(shrinkage method)을 활용하여 다중공선성 문제를 해소 할 수 있다. 대표적인 차원축소방법으로 주성분 회귀분석(principal component regression)을 이용하여 설명변수들의 본래 성질을 유지하되, 변환된 소수의 상위주성분으로 회귀분석을 수행한다. 그러나, 선택된 상위주성분이 반응변수와 연관성이 적을 경우 예측력이 떨어지고 추정된 설명변수의 선형결합으로 변환된 주성분을 해석하는 데에도 어려움이 있다. 따라서, 본 논문은 희소벌점함수(sparse-inducing penalty function)를 적용한 축소추정법을 통해 추정량의 크기를 축소시키는 동시에 반응변수와 연관성이 높은 몇개의 주성분만을 선택하여 예측력과 설명력을 높이는

방안을 제시한다.

본 논문은 총 5장으로 구성되어 있다. 제1장은 서론으로서 연구배경 및 목적과 연구방법을 서술하였다. 제2장은 주성분 로지스틱 회귀분석에 대해 자세하게 설명하였고, 제3장은 벌점을 부여한 주성분 로지스틱 회귀분석에 대해 서술하였다. 제4장에서는 모의실험을 통하여 제안된 방법의 효율성에 대해 알아볼 수 있는 모의실험 설계와 결과를 수록하였다. 제5장은 결론으로 본 연구의 결과를 요약하였다.

2 주성분 로지스틱 회귀분석법

2.1 로지스틱 회귀분석(Logistic Regression)

반응변수 Y가 0 또는 1의 값을 갖는 이변량 질적변수(binary qualitative variable)일 때, 설명변수와 반응변수 간의 함수 관계를 로지스틱 회귀모형으로 적합한다. 이분형 자료가 관심의 범주인 $y = 1$에 속할 확률이 $\Pr(Y = 1) = \theta$, 그 반대의 경우를 $\Pr(Y = 0) = 1 - \theta$ 이라고 하자. 이때, 설명변수 $X = (x_1, x_2, \ldots, x_p)$에 대한 베르누이 분포(Bernoulli distribution)를 따르는 반응변수 Y의 조건부 기댓값을 다음과 같이 표현할 수 있다.

$$Y \sim \text{Bernoulli}(\theta),$$

$$\Pr(Y = y|X = \mathbf{x}) = [\theta(\mathbf{x})]^y[(1 - \theta(\mathbf{x}))]^{1-y} \qquad y \in \{1, 0\},$$

$$E(Y|X=\mathbf{x}) = \Pr(Y=1|X=\mathbf{x}) = \theta(\mathbf{x}).$$

반응변수의 조건부 기댓값은 성공확률이므로 $0 \le \theta(x) \le 1$ 의 제약을 가지게 되는데, 이런 경우 일반적인 선형회귀 모형을 적용하게 되면 확률값이 가지는 범위를 벗어나게 된다. 이러한 문제를 해결하기 위해 확률의 로짓변환(logit transformation)을 이용하여 선형화 한다.

$$\log(\theta(\mathbf{x})/(1-\theta(\mathbf{x}))) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

위 식을 정리하면 다음과 같고, 이는 0과 1사이의 값을 가지는 확률의 성질을 만족하게 된다.

$$\theta(\mathbf{x}) = (\exp(\beta_0 + \mathbf{x}^T\boldsymbol{\beta}))/(1+\exp(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})) = \{1+\exp(-\beta_0 - \mathbf{x}^T\boldsymbol{\beta})\}^{-1}.$$

로지스틱 회귀모형에서 모수를 추정할 때, 일반적으로 최대우도추정법(maximum likelihood estimation)을 이용한다. n쌍의 $(\mathbf{x}_i,\, y_i)$, $i = 1, \cdots, n$ 의 자료를 가정하자. 이때, 아래와 같이 모수에 대한 우도함수(likelihood function)를 정의하고, 우도함수가 최대화 되는 모수를 추정치로 구한다.

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \prod_{i=1}^{n} \Pr(Y=y_i|X=\mathbf{x}_i) = \prod_{i=1}^{n} \theta(\mathbf{x}_i)^{y_i}(1-\theta(\mathbf{x}_i))^{1-y_i}.$$

계산의 편의를 위해 위의 우도함수에 자연로그를 취해서 계산한다.

$$\begin{aligned}\log L(\beta_0, \beta_1, \ldots, \beta_p) &= \sum_{i=1}^{n} \log(\Pr(Y=y_i|X=\mathbf{x}_i)) \\ &= \sum_{i=1}^{n} [y_i(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) + \log\{1+\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})\}].\end{aligned}$$

로그우도함수식이 최대화 되는 모수 $\beta_0, \boldsymbol{\beta}$를 구하기 위해 $\beta_0, \boldsymbol{\beta}$ 각각에 대해 편미분하여 0이 되는 값을 구한다. 다음을 구할 때는 Newton-Raphson 방법 등을 이용한 반복 알고리즘을 통해 수치적(numerical)으로 추정한다.

$$U(\beta_0) = \frac{\partial L(\beta_0, \beta_1, \ldots, \beta_p)}{\partial \beta_0} = 0,$$
$$U(\boldsymbol{\beta}) = \frac{\partial L(\beta_0, \beta_1, \ldots, \beta_p)}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

2.1.1 주성분 로지스틱회귀분석(Principal Component Logistic Regression)

회귀분석에서 설명변수들 사이에 높은 상관관계가 존재하면 다중공선성(multicollinearity)을 의심해볼 수 있다. 다중공선성이 존재하는 경우, 추정된 회귀계수의 분산이 커지므로 불안정한 회귀계수 추정값이 도출된다. 일반적으로 차원축소방법 중의 하나인 주성분분석(Principal Component Analysis)을 통해 다중공선성을 해결한다. 주성분분석은 고차원 데이터의 변동을 변수들간의 선형결합(linear combination)을 통해 나타내고, 이들 중 일부의 주성분을 선택하여 저차원 공간으로 자료를 표현한다. 즉, 자료의 큰 손실 없이 몇 개의 주성분을 선택하여 원자료 전체의 변동을 설명하므로 차원이 축소되는 효과를 가져올 수 있으며, 각 주성분들이 서로 직교하기 때문에 다중공선성의 문제도 해결 할 수 있다.

주성분로지스틱회귀법(PCLR)은 로지스틱 회귀분석에서 설명변수들

간에 강한 상관관계가 있을 때 해결하는 대표적인 방법이다. 설명변수의 분산을 최대로 하는 선형결합을 통하여 주성분(principal component)을 순차적으로 생성하고, 설명변수의 특성은 최대한 유지하면서 자료의 변동을 가장 잘 설명하는 상위 주성분들을 이용하여 회귀분석을 수행하므로 다중공선성의 문제를 해결할 수 있다. 주성분의 일부만 이용함으로써 주성분 회귀모형을 통해 얻은 주성분 회귀계수는 편의추정량(biased estimator)이 되지만, 추정량의 분산을 줄이는 효과가 있어서 예측력을 향상시킨다는 점에서 장점이 있다.

n개의 자료와 p개의 설명변수로 이루어진 자료행렬 $\mathbf{X} = (x_{ij})_{i=1,\ldots,n,j=1,\ldots,p}$을 가정하자. $\mathbf{X}$는 $n \times p$ 크기의 행렬이고, 중심화(centering) 되어있다고 하자. 또한, 반응변수인 $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ 도 중심화 되어있다고 가정하자. 설명변수의 공분산행렬(covariance matrix)은 $\mathbf{A} = \mathbf{X}^T\mathbf{X}/(n-1)$로 표현된다. 공분산행렬에 대한 고유벡터(eigen vector)를 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$라 하면 각 고유벡터 $\mathbf{v}_k$에 대하여 대응되는 주성분(principal component)은 다음과 같이 얻을 수 있다.

$$\mathbf{z}_k = \mathbf{X}\mathbf{v}_k, \qquad k = 1, 2, \ldots, p.$$

크기가 n인 벡터 $\mathbf{z}_k = (z_{1k}, z_{2k}, \ldots, z_{nk})^T$는 p차원 공간 상의 점으로 표현되는 n개의 자료의 설명변수를 고유벡터 $\mathbf{v}_k$방향으로 직교사영(orthogonal projection)된 점의 좌표이다. 주성분분석에서 $\mathbf{v}_k$를 k번째 주성분방향(PC loading)이라 부르며, 해당 $\mathbf{z}_k$를 k번째 주성분이라 부른다. 즉, i번째 자료의 k번째 주성분은 $z_{ik} = \mathbf{x}_i^T\mathbf{v}_k$로 주어지게 되며, 여기서 $\mathbf{x}_i =$

$(x_{i1}, x_{i2}, \ldots, x_{ip})^T$는 $\mathbf{X}$의 i번째 행이다. 전체 주성분은 다음의 행렬계산으로 쉽게 얻을 수 있다.

$$\mathbf{Z} = (z_{ik}) = \mathbf{XV}.$$

위 식에서 $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p)$ 이다. 또한, $\mathbf{Z}^T\mathbf{Z} = \mathbf{V}^T\mathbf{X}^T\mathbf{XV} = (n-1)\mathbf{\Lambda}^T\mathbf{A}\mathbf{\Lambda} = (n-1)\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V} = (n-1)\mathbf{\Lambda}$ 로 주어짐을 확인할 수 있다. 중심화 되어있는 설명변수와 반응변수로 이루어진 자료에 대한 선형회귀모형은 절편(intercept)이 없는 아래의 모형으로 표현이 가능하며, 이를 주성분 행렬($\mathbf{Z}$)로 표현할 수 있다.

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}.\end{aligned}$$

위의 식에서 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$이고 $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$이다. 즉, 설명변수 대신 주성분을 설명변수로 하는 회귀분석의 결과는 본래의 설명변수로 표현된 회귀모형과 동일함을 알 수 있다. 변경된 회귀식의 모수 $\boldsymbol{\gamma}$의 추정치는 다음과 같이 주어진다.

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}.$$

위 식을 이용하여 본래의 회귀계수의 추정치는 다음 식을 통해 얻을 수 있다.

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}.$$

주성분회귀모형은 설명변수의 변동의 대부분을 설명하는 주요 주성분을 사용하여 회귀분석을 수행하는 방법이다. 즉, 전체 p개의 주성분을 이용

하는 대신, 상위 K개의 주성분 만을 설명변수로 고려한다. 이를 이해하기 위해서 $\hat{\beta}$의 분산을 계산하면 다음과 같다.

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T = \sigma^2\sum_{k=1}^{p}\frac{1}{\lambda_k}\mathbf{v}_k\mathbf{v}_k^T.$$

하위 고유치값(λ_k)이 0에 가깝다면 추정량의 분산이 매우 커지게 됨을 알 수 있다. 따라서, 고유치의 크기가 0 에 가까운 하위 주성분을 제거한다면 추정치의 분산을 줄일 수 있다. 이처럼, 주성분회귀분석법은 하위 주성분을 제거하면서 추정량의 분산을 줄여주고, 고차원의 자료를 저차원 상에서 분석 할 수 있다는 장점을 가지고 있다. 그러나 차원을 줄인다고 해도 실질적으로 본래의 설명변수들을 모두 가지고 있으므로 해석에 용이하지 않는 단점을 가지고 있다.

3 벌점함수를 부여한 주성분 로지스틱 회귀분석

3.1 벌점회귀모형(Penalized regression)

일반 로지스틱 회귀분석에서는 최대우도추정법(Maximum likelihood estimation)을 이용하여 로그우도함수(log-likelihood function)를 최대로 하는 회귀계수를 추정한다. 그러나 자료의 수보다 설명변수의 수가 더 많아지게 되면, 여러 가지 문제점이 발생한다. 최대우도추정법에 의해 추정된 추정값이 유일하지 않고, 추정량의 분산이 커져 신뢰성을 잃게 된다. 또한,

설명변수가 많아서 해석상 어려움이 생기며, 서로 상관관계가 있는 설명변수가 있을 가능성이 높아지게 되어 다중공선성 문제가 발생할 가능성이 높다. 따라서, 이런 단점을 보완할 수 있는 해결책으로서 최대우도추정법에 벌점(penalty)을 부여하여 회귀계수를 축소하는 Ridge regression이나 Lasso regression 방법을 생각해 볼 수 있다.

3.1.1 능형 회귀분석법(Ridge regression)

능형회귀분석(ridge regression)은 회귀계수 추정량에 약간의 편차(bias)를 허용하는 대신, 분산을 감소시켜 추정량의 신뢰성을 높이는 방법이다. 뿐만 아니라 예측력을 개선시켜주어 데이터에 잘 맞는 모형을 도출할 수 있다. 다음과 같이 반응변수가 이항자료 일 때, 로지스틱 회귀모형을 가정하자.

$$y_i = \theta_i + \epsilon_i. \qquad (i = 1, \ldots, n)$$

로그우도함수(log-likelihood function)를 최대로 하는 최대우도추정량을 이용하여 최적의 모수를 구할 수 있다.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\arg\min} \; -2\log L(\beta_0, \beta_1, \ldots, \beta_p)$$

능형회귀추정량은 다음과 같이 특정 L_2 제약조건 하에서 로그우도함수를 최대로 하는 모수값이다.

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\arg\min} \; -2\log L(\beta_0, \beta_1, \ldots, \beta_p) \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t.$$

위 식에서 t는 양수로써 모수의 크기를 조절하는 조절모수(tuning parameter)이다. t가 충분히 크면 모수에 대한 제약이 없기 때문에 최대우도 추정량과 일치한다. 반면, 조절모수 t가 0이면 회귀계수가 모두 0이 된다. 조절모수 t가 0은 아니지만 0에 가까이 작아지면 최대우도 추정량의 크기가 작아지는 효과를 가져올 수 있다.

3.1.2 Lasso 회귀분석법(Lasso regression)

Lasso 회귀분석법은 능형회귀분석법(Ridge)과 같이 회귀계수를 추정할 때, 적절한 제약조건하에서 최적의 모수를 추정하는 방법이다. Tibshirani (1996)에 의해 고안된 Lasso 회귀기법은 회귀계수 추정량에 약간의 편의(bias)를 주는 대신 분산을 감소시키면서 추정량의 예측도를 높일 수 있다. 뿐만 아니라, 능형회귀분석과 다르게 L_1 제약조건 하에 유의하지 않은 회귀계수를 정확하게 0으로 보내면서 자동으로 변수선택(variable selection)의 효과를 준다. 따라서, 추정량의 예측도 향상 및 자동 변수선택의 효과로 인해 많은 분야에서 활발하게 활용되고 있다. 다음은 Lasso 회귀추정량을 산출하기 위한 특정 제약조건하에서의 최대우도추정량 식이다.

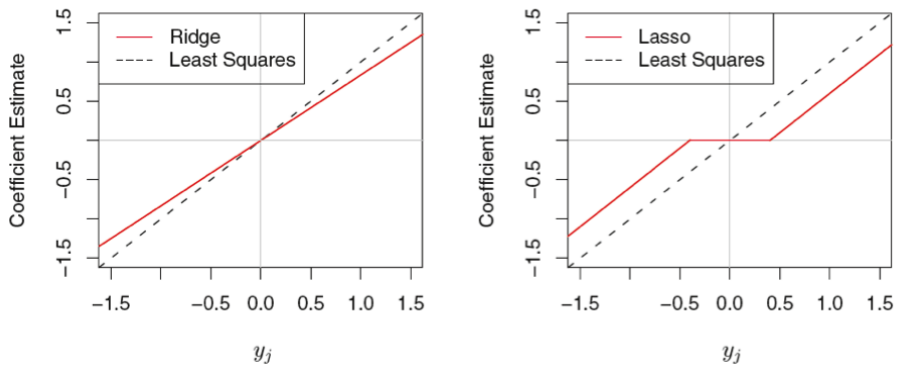
$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\arg\min} \sum_{i=1}^{n} -2\log L(\beta_0,\beta_1,\ldots,\beta_p), \quad \text{subject to} \sum_{j=1}^{p} |\beta_j| \le t.$$

마찬가지로, t는 양수값을 가지는 조절모수이다. t값이 충분히 크면 모수에 대한 제약이 없으므로 최대우도추정량(maximum likelihood estimate)과 같은 값을 갖는다. 반면에 t값이 0이면, 모든 추정치가 0이 된다. 따라

서, 적절하게 t의 크기가 0에 가까워지면 회귀계수 추정치가 작아지게 되고, 일부분은 정확히 0으로 주어지게 된다. 결과적으로 Lasso회귀모형은 통계적으로 유의하지 않은 몇 개의 회귀계수가 0 이 되면서 변수가 축소되고, 축소된 변수들을 기반으로 추정된 회귀계수는 편의(bias)가 있으나 분산(variance)은 매우 작아지게 되어 추정량의 정확도와 예측도가 높아진다. 또한 변수가 축소되면서 해석이 용이하게 되는 장점도 있다.

3.1.3 능형회귀와 Lasso회귀의 비교

그림 1: 최소제곱회귀추정치와 능형회귀 및 lasso회귀 추정치의 비교그림.

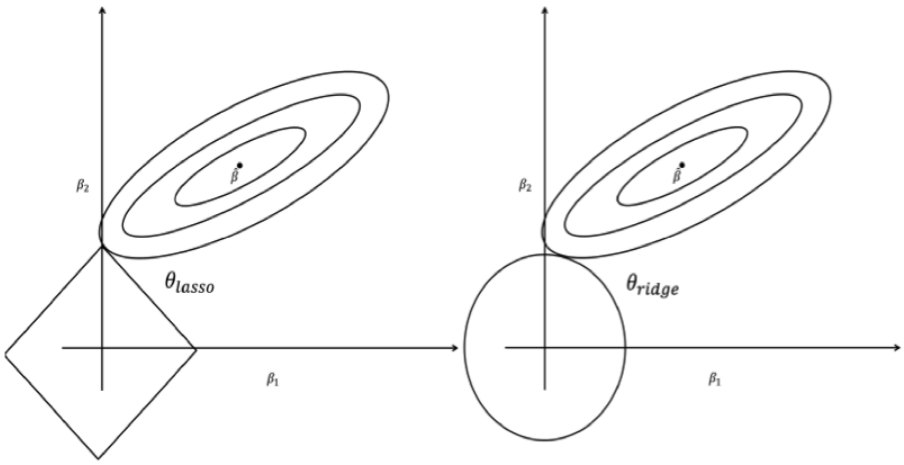


능형회귀방법과 Lasso회귀방법을 통해 회귀계수를 추정할 때, 최대우도추정법에 제약조건이 주어진다. 능형회귀의 경우 L_2-penalty, Lasso회귀

는 L_1-penalty의 제약조건하에 추정량을 구하게 된다. 우선, 두 추정량을 비교하기 위해 $p = 2$인 경우를 가정한다. 그림 1은 최대우도추정치에 따른 Ridge와 Lasso회귀 추정치의 변화를 나타낸 것이다. 그림의 좌우 그래프에서 붉은색 실선은 능형회귀 및 lasso회귀의 추정치이며 검은색 점선은 최대우도추정치와의 비교를 위한 참조선이다. 최대우도추정치에 비해 두 추정치가 일괄적으로 작게 추정되며 최대우도추정치가 일정수준 이하의 값을 가지게 되면 Lasso 회귀 추정치는 정확히 0으로 주어지는 것을 볼 수 있다. 이는 능형회귀분석은 회귀계수 추정치를 축소시키는 역할을 하지만 lasso 회귀분석은 추정치 값을 정확히 0으로 만들어 변수선택의 역할을 하는 것을 알 수 있다.

그림 2는 차례로 lasso회귀모형의 L_1 제약식 $|\beta_1| + |\beta_2| \leq t$와 능형회귀모형의 L_2 제약식 $\beta_1^2 + \beta_2^2 \leq t$의 영역 하에서 회귀계수 추정량을 얻는 과정에 대한 기하학적 묘사이다. 두 방법은 모두 동일한 로그우도함수 $(-2 \log L(\beta_0, \cdots, \beta_p))$를 가지고 있으며 이는 그림상에서 타원 형태의 등고선으로 되어있다. 그림에 등고선의 가장 낮은 위치에 표현된 점을 최대우도추정치($\hat{\beta}$)에 대응된다. 각 방법의 추정치는 제약식 영역 하에서 로그우도함수값을 최대로 하는 값이다. lasso회귀모형의 경우 제약식 $|\beta_1| + |\beta_2| \leq t$의 영역은 마름모 형태로 표현이 되고, 능형회귀분석의 경우 제약식 $\beta_1^2 + \beta_2^2 \leq t$이 원형으로 표현된다. 이 제약영역의 점 중에서 등고선과 최초로 만나게 되는 점이 각 방법에 대응하는 회귀계수 추정치가 된다. 따라서 lasso회귀분석의 제약영역이 마름모형태를 가지면서, 마름모의 각 꼭지점이 등고선과

그림 2: Lasso회귀법 및 능형회귀법의 기하학적 형태의 비교그림.



닿을 때 축의 일부분 값이 0에 해당되므로 lasso회귀계수 추정량은 일부분이 0이 될 가능성이 높다. 반면에 능형회귀분석의 경우는 원형의 제약식을 가지므로, 축 위에 놓인 점이 회귀계수가 될 가능성이 거의 없다.

3.2 벌점회귀에 기반한 주성분 로지스틱 회귀분석법

주성분 회귀에서는 상위주성분을 설명변수로 하여 차원축소를 통해 추정량의 성능을 개선하지만, 반응변수와 연관성이 높은 하위주성분을 활용하

지 않음으로 인하여 효율성이 떨어지는 단점을 지적하였다. 이를 개선하는 직관적인 방법은 주성분을 설명변수로 사용하되 반응변수와 연관성이 높은 주성분을 선택하여 이를 회귀모형에 활용하는 것이다. 본 연구에서는 주성분을 선택하는 방법으로써 Lasso, Ridge, SCAD, Adaptive lasso 등의 벌점함수를 활용하여 주성분을 자동으로 선별하는 방법을 제안한다. 이를 통해 반응변수에 유의한 영향을 주는 주성분은 하위주성분이라 하더라도 모형에 들어감으로써 설명력의 감소를 방지할 수 있고, 상위주성분이라 하더라도 반응변수에 대한 설명력이 작으면 벌점최적화에 의해 최종모형에서 배제함으로써 최적의 주성분 집합을 구성한다.

크기가 $n \times p$인 $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)^T$ 는 n개의 자료로부터 얻은 p개의 설명변수로 구성된 자료행렬이라 하자. 즉, $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$는 i번째 관측치의 설명변수의 모임으로 해석할 수 있다. 이를 이용하여 다중선형회귀모형을 다음과 같이 표현할 수 있다.

$$\log(\theta_i/(1-\theta_i)) = \mathbf{x}_i^T \boldsymbol{\beta}_i.$$

설명변수를 주성분으로 변환하면

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

로 표현할 수 있으며, 변환한 새로운 변수를 이용하여 다음과 같은 모형을 얻을 수 있다.

$$\log(\boldsymbol{\theta}_i/(1-\boldsymbol{\theta}_i)) = \mathbf{x}_i^T \boldsymbol{\beta} = \mathbf{Z}\mathbf{V}^T\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\gamma}.$$

여기서 $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$이다. 행렬 $\mathbf{Z}$를 구성하는 열벡터 $\tilde{\mathbf{z}}_j$ $(j = 1, 2, \cdots, p)$는 j번째 주성분으로 주어진다. 회귀분석 시 이 주성분의 전부를 설명변수로 삼고 반응변수에 회귀적합을 한다면 주성분 $\mathbf{Z}$에 대한 회귀계수 $\boldsymbol{\gamma}$를 추정할 수 있다. 그 추정치를 원래 모형의 회귀계수 $\boldsymbol{\beta}$로 역변환하여 추정량을 구하면 정확히 최대우도추정량이 된다. $\boldsymbol{\gamma} = \mathbf{V}^T\boldsymbol{\beta}$의 관계로부터 $\hat{\boldsymbol{\beta}}$의 최대우도추정량은

$$\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$$

임을 알 수 있다. 여기서 모수 $\boldsymbol{\beta}$에 대하여 추정할 때 모든 주성분을 사용하여 $\boldsymbol{\gamma}$를 추정한다면 최대우도추정량과 동일한 결과를 얻는다.

일반적인 주성분회귀분석은 작은 고유치와 대응되는 몇 개의 주성분을 제거시켜 차원을 축소한다. 하지만 이러한 주성분의 선택은 작은 고유치에 대한 기준이 존재하지 않아 주관적인 결정을 해야만 하는 단점이 있으며, 하위주성분의 일관적인 배제로 인하여 효율성이 떨어질 우려가 있다. 따라서 이러한 주성분 선택의 방법을 본 논문에서는 벌점함수의 도입으로 선택하고자 한다. 이러한 분석법을 본 연구에서는 벌점 주성분 로지스틱회귀분석 이라고 부른다. 본 연구에서는 벌점함수로써 능형벌점 (ridge penalty), lasso벌점 (lasso penalty), SCAD벌점 (SCAD penalty)등을 이용한다. 벌점함수의 형태는 다음과 같다.

- 능형벌점함수

$$\sum_{j=1}^{p} \beta_j^2 \leq t$$

- Lasso벌점함수

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

- SCAD벌점함수

$$\sum_{j=1}^{p} P_\lambda |\beta_j| \leq t$$

- Adaptive lasso 벌점함수

$$\sum_{j=1}^{p} W_j |\beta_j| \leq t$$

4 모의실험

본 장에서는 3장에서 소개한 벌점 로지스틱 회귀에 기반한 주성분의 선별 방법을 모의실험에 적용하고 각각의 분류법들의 성능을 비교한다.

4.1 모의실험 설계

본 논문에서는 통계프로그램 R을 사용하여 다음과 같이 모의실험 상황을 설정하였다. 자료의 수를 n, 변수의 개수가 p인 설명변수 행렬을 $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_p^T)^T$라 하고, 반응변수 행렬을 $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T \in \{1, 0\}^n$라 하자. 설명변수와 반응변수의 관계는 다음과 같은 선형식을 가정한다.

$$\log(\theta_i/(1-\theta_i)) = \mathbf{x}_i^T \boldsymbol{\beta}_i. \qquad (i = 1, \cdots, n)$$

컴퓨터 모의실험을 위하여 설명변수 $\mathbf{X}$, 회귀계수 $\boldsymbol{\beta}$ 그리고 반응변수 $\mathbf{y}$를 생성하기 위해 다음 작업을 수행한다. 먼저, 실험자료를 생성하기 위해 다음 식의 $\boldsymbol{\gamma}$값을 설정한다.

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\gamma}.$$

설명변수의 선형결합을 통해 얻어낸 주성분 회귀분석에서, γ_j $(j = 1, 2, \ldots, p)$의 값은 j번째 주성분과 반응변수의 연관성을 나타내며 $\gamma_j = 0$은 j번째 주성분이 반응변수에 대한 설명력이 전혀 없음을 의미한다. 주성분과 반응변수 간의 관계를 아래와 같이 모형화 한다.

(T1) $\boldsymbol{\gamma} = (1, 1, 1, 0, 0, 0, \ldots, 0, 0)^T$

(T2) $\boldsymbol{\gamma} = (0, \ldots, 0, 1, 0, 0, 1, 0, 1)^T$

(T3) 랜덤설정

(T1)의 경우, 상위 3개의 주성분만이 반응변수에 동일한 설명력을 가지는 경우이다. (T2)의 경우는 끝에서 1, 3, 6번째 주성분만이 반응변수에 동일한 설명력을 가진다.(T3)는 전체 주성분 중에 랜덤하게 3개만을 선택하여 반응변수에 동일한 설명력을 가진다. 다음 3가지 상황에 대해서 각 자료를 동등하게 생성한다.

(Step1) $\mathbf{U} = (\mathbf{u}_1^T, \mathbf{u}_2^T, \ldots, \mathbf{u}_{1000}^T)^T = (u_{ij}) \in \mathbb{R}^{1000 \times p}$의 생성한다.

$1000 \times p$개의 확률변수 $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{1000p}$를 $N(0, 1)$로부터 랜덤하게 생성하여 $\mathbf{U}$를 만든다.

(Step2) 1000개의 샘플에 대한 표본공분산행렬을 계산한다.

$\mathbf{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{1000}(\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T$를 계산한다. 여기서, $\bar{\mathbf{u}} = \frac{1}{1000}\sum_{i=1}^{1000}\mathbf{u}_i$ 이다.

(Step3) 고유치분해(eigen decomposition)을 이용하여 직교행렬 $\mathbf{V}$를 구한다.

$\mathbf{\Sigma}^* = \mathbf{V}\mathbf{D}\mathbf{V}^T$를 계산하여 $\mathbf{V}$를 이어지는 계산에 활용한다. 여기서 $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_p)$이고, d_l은 l번째 고유치이다. $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p)$ 이고, $\mathbf{v}_l$은 해당 고유벡터이다.

(Step4) n개의 자료에 대한 주성분점수를 생성한다.

n개의 주성분 $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$을 $N(0, \mathbf{\Lambda})$로부터 랜덤하게 생성하여 주성분행렬 $\mathbf{Z} = (\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_n^T)^T$를 구성한다. 여기서 $\mathbf{\Lambda}$는 $\lambda_j = 1.2^{1-j}$ $(j = 1, 2, \ldots, p)$를 원소로 하는 대각행렬로 정의한다. λ_j는 j번째 주성분의 분산이다.

(Step5) $\mathbf{X} = \mathbf{Z}\mathbf{V}^T$의 변환을 통해 설명변수를 생성한다.

$\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)^T$이며 $\mathbf{x}_i = \mathbf{V}\mathbf{z}_i$ $(i = 1, 2, \ldots, n)$이다.

(Step6) $\mathbf{Z}$를 이용하여 반응변수 $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$을 생성한다.

미리 설정한 $\boldsymbol{\gamma}$를 이용하여 확률값을 생성한다.

$$p_i = \frac{\exp(\mathbf{z}_i^T\boldsymbol{\gamma}_i + \beta_0)}{(1 + \exp(\mathbf{z}_i^T\boldsymbol{\gamma}_i + \beta_0))}$$

$n \times p$개의 확률변수 $y_1, y_2, \ldots, y_n$를 Binomial(n, p_i) 분포로부터 랜덤하게 생성하여 $\mathbf{y}$를 만든다.

위에서 반응변수 $\mathbf{y}$는 $\mathbf{Z}$와 $\boldsymbol{\gamma}$에 의해 생성된다. 이와 같은 정보를 갖는 설명변수 $\mathbf{X}$이고 해당 회귀계수는 $\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\gamma}$로 주어진다. 절편 $\beta_0 = 0$으로 가정한다. 컴퓨터 모의실험에서는 주성분 및 주성분에 대응하는 회귀계수 $\boldsymbol{\gamma}$를 알고 있지만, 일반적인 회귀분석을 수행하게 되는 경우 설명변수 $\mathbf{X}$ 및 반응변수 $\mathbf{y}$만 주어지고 이를 기반으로 회귀분석을 수행하게 된다. 위 과정은 $\mathbf{x}_i$들은 공분산 $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$이고 평균이 $\mathbf{0}$인 다변수정규분포에서 생성한 것과 동일함을 알 수 있다.

자료의 개수(n)과 변수의 개수(p)는 다음과 같이 설정하였다.

- 자료의 개수가 변수의 개수보다 많은 경우: $p = 10$, $n = 100, 500, 1000$
- 변수의 개수가 자료의 개수보다 많은 경우: $n = 100$, $p = 50, 100, 150$

위와 같이 생성한 자료를 기반으로 본 연구에서 제안한 방법과 다양한 회귀분석 방법을 적용하여 성능을 비교한다. 비교하고자 하는 방법은 다음과 같다.

(M1) PPCLR-L : 주성분을 설명변수로 하고, lasso penalty를 이용하여 주성분을 선택하여 회귀계수 추정

(M2) PPCLR-R : 주성분을 설명변수로 하고, ridge penalty를 이용하여 주성분을 선택하여 회귀계수 추정

(M3) PPCLR-S : 주성분을 설명변수로 하고, scad penalty를 이용하여 주성분을 선택하여 회귀계수 추정

(M4) PPCLR-A : 주성분을 설명변수로 하고, adaptive lasso penalty를 이용하여 주성분을 선택하여 회귀계수 추정

(M5) PLSLR : 설명변수를 부분최소제곱분석으로 회귀계수 추정

(M6) PCLR : 원변수를 설명변수로 하는 주성분 로지스틱 회귀분석

(M7) PLR-L : 원변수를 설명변수로 하고, lasso penalty를 이용하여 회귀계수 추정

(M8) PLR-R : 원변수를 설명변수로 하고, ridge penalty를 이용하여 회귀계수 추정

(M9) PLR-S : 원변수를 설명변수로 하고, scad penalty를 이용하여 회귀계수 추정

(M10) PLR-A : 원변수를 설명변수로 하고, adaptive lasso penalty를 이용하여 회귀계수 추정

방법 (M1)~(M4)는 본 연구에서 제안한 방법으로서 주성분을 설명변수로 사용한다. 여기서 주성분은 (Step 4)에서 자료를 생성하기 위해 준비한 $\mathbf{Z}$를 사용하지 않고 설명변수 행렬 $\mathbf{X}$의 주성분 변환을 통해 얻은 주성분을 사용하였다. 그 이유는 실제자료가 주어졌을 때 주성분을 추정해서 사용해야 하는 현실적인 상황을 반영하기 위함이다. 고려하는 방법 중 PLSLR을 제외한 모든 방법은 각기 모형선택(model selection)을 위해 조절모수를

포함하고 있다. 이를 10-fold cross validation(CV)을 통해 최적의 모수를 선택하였다.

위의 각 방법을 통해 얻은 회귀계수추정량을 $\hat{\boldsymbol{\beta}}$라 하고, 각 방법의 성능을 평가하기 위해 추정오차(estimation error)를 다음과 같이 계산하였다.

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = \sqrt{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}.$$

여기서, 추정오차가 작다는 것은 추정된 회귀계수 $\hat{\boldsymbol{\beta}}$가 실제 $\boldsymbol{\beta}$와 큰 차이가 없으며 이는 회귀계수를 잘 추정했다는 것을 의미한다. 모의실험을 반복할 때마다 비교결과가 항상 동일하지 않으므로 위의 실험을 반복하여 수행하였다. 각 상황에 대하여 1,000회의 모의실험을 반복하고 각 결과로부터 얻어진 1,000개의 추정오차의 평균값을 구하여 각 모형의 성능을 비교하였다.

4.1.1 모의실험 결과

앞서 언급한 방법론에 대하여 모의실험 결과를 얻었다. 모의실험 설계에서 제안한 세가지 상황(T1 T3)에 따라, 자료를 1000회 반복 생성하여 얻은 회귀계수 추정량의 표준오차 값을 비교할 수 있게 표1 표6에 수록하였다. 표1 표3은 설명변수의 수가 $p = 10$으로 고정되어, 자료의 수가 ($n = 100, 500, 1000$) 증가함에 따라 제안된 각 상황별 추정오차 평균값을 나타낸다. 표4 표6은 자료의 수가 $n = 100$으로 고정되어 있을 때, 설명변수의 수가 ($p = 50, 100, 150$) 증가함에 따라 상황별 추정오차의 평균값을 나타낸다.

표1은 T1에서 고정된 설명변수($p = 10$)에 대해 자료의 크기를 증가시키면서 얻은 결과이다. 자료의 크기가 커짐에 따라 추정오차의 평균값이 작아지는 경향을 확인할 수 있다. 또한, 주성분을 각 희소벌점함수에 도입한 경우가 본래의 설명변수를 도입한 경우에 비해 낮은 추정오차의 평균값을 가지는 것을 알 수 있다. 이를 통해, 선형 변환된 주성분이 일반 설명변수보다 회귀계수를 더 정확하게 추정한다고 할 수 있다. 따라서, 본 논문에서 주장하고 있는 주성분을 기반으로 한 벌점 로지스틱 회귀분석법이 더 정확한 추정결과를 보인다고 할 수 있다. 표2는 T2에서 고정된 설명변수($p = 10$)에 대해 표1과 같은 방식으로 자료의 크기를 증가시키면서 얻은 결과이다. 자료의 크기가 커짐에 따라 추정오차의 평균값이 대체적으로 작아지는 경향을 보이고 있다. 또한, 주성분에 각 희소벌점함수를 도입하여 산출한 결과 값이 원자료인 설명변수에 희소벌점함수를 도입하여 얻은 결과 값보다 더 낮게 나타난다. 반면, 모의실험을 할 때 이용된 통계 소프트웨어 R에서 $n = 1000$일 때 계산량이 증가하여 원자료 설명변수에 SCAD 벌점함수를 도입한 결과(PLR-S)를 얻을 수 없었다. 그러나 주성분에 SCAD 벌점함수를 도입하여 얻은 결과가(PPCLR-S) 추정오차의 평균값이 가장 낮게 나온 것을 미루어 볼 때, PLR-S의 결과도 나쁘지 않았음을 예상 할 수 있다. 표3의 경우 T3일 때의 표1, 표2와 같은 방식으로 얻은 결과이다. 각 자료의 크기에서 본 연구가 제안하는 주성분에 희소벌점함수를 도입한 방법이 원자료 설명변수를 기반으로 벌점함수를 도입한 것 보다 더 낮은 추정오차의 평균값을 도출하였다. 마찬가지로 $n = 1000$일 때의 PPCLR-S

는 계산상의 어려움이 있어 결과를 얻지 못하였다.

표4는 T1에서 고정된 자료의 수($n = 100$)에 대해 설명변수의 크기를 증가시키면서 얻은 결과이다. 각 설명변수의 크기에 대하여 주성분을 기반으로 하여 벌점함수를 도입한 결과가 원자료 설명변수를 기반으로 벌점함수를 도입한 것 보다 추정오차의 평균값이 더 낮게 나왔다. 이는 설명변수가 증가하는 상황에서도 본 연구가 제안한 주성분에 벌점함수를 도입한 방법들이 일반 주성분 로지스틱 회귀분석보다 더 정확하게 추정함을 알 수 있다. 표5는 T2에서 고정된 자료의 수($n = 100$)에 대해 설명변수의 크기를 증가시키면서 얻은 결과이다. 통계 소프트웨어의 계산적 문제로 인해 $p = 100, 150$의 경우 PLR-S, PPCLR-S의 결과 값을 얻을 수 없었다. 그러나, 각 설명변수의 크기에서 보면 본 논문에서 제안하는 주성분을 기반으로 희소벌점함수를 도입한 방법의 결과가 더 좋음을 알 수 있다. 표6은 T3에서 표4, 표5와 같은 방식으로 얻은 결과이다. 각 설명변수의 크기에서 본 논문에서 제안하는 주성분을 기반으로 희소벌점함수를 도입한 결과들이 더 정확한 추정을 한다고 볼 수 있다. 마찬가지로, 표5와 동일한 이유로 $n = 100, 150$ 일 때의 PLR-S, PPCLR-S의 결과를 얻지 못했다.

5 결론

본 연구를 통해, 주성분을 활용한 로지스틱 회귀모형에 벌점함수를 도입하여 새로운 회귀법을 제시하였다. 모의실험을 수행하여 얻은 결론을 토대로

제시한 방법론의 성능이 어느 정도 입증됨을 확인 할 수 있다. 일반적으로 반응변수가 이변량 질적변수인 고차원 데이터를 로지스틱 모형에 적합시키는 경우, 다중공선성이 존재할 가능성이 높아 왜곡된 결과를 낳게 된다. 변수간의 강한 상관성으로 발생하는 다중공선성 문제를 해결하면서 좀 더 정확한 모델추정을 위한 해결책으로 본 논문에서 제안하는 주성분을 기반으로 벌점함수를 적용한 방법은 차원축소법에 변수선택의 개념으로 확장시킬 수 있으며, 그 결과가 매우 효과적이라 할 수 있다. 실제 데이터를 이용하여 실증적 실험을 통한 모델 추정으로 검증하는 것이 추후 연구과제이다.

참고문헌

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hoerl, A. E. and Kennard, R. W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A. E. and Kennard, R. W. (1970b) Ridge regression: Iterative estimation of the baising parameter. *Technometrics*, **12**, 77–88.

James, G., Witten, D., Hastie, T., and Tobshirani, R. (2013) *An Introduc-*

tion to Statistical Learning: Linear Model Selection and Regularization. Springer.

Jolliffe, I. T. (2002) *Principal Component Analysis* . Springer.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B.* **58**, 267–288.

The Lasso Logistic Regression Model: Modifications to aid causality assessment for Adverse Events Following Immunization (2010)

6 부록

6.1 표

표 1: type 1에서 $p = 10$이고 n이 증가할 때 추정오차의 평균.

gam=1	$n = 100$	$n = 500$	$n = 1000$
PLSLR	0.3688	0.1942	0.1534
PLR-A	0.3874	0.1703	0.1198
PLR-L	0.3891	0.1714	0.1198
PLR-R	1.3471	0.2584	1.3229
PLR-S	0.3787	0.1703	0.1195
PPCLR-A	0.3874	0.1703	0.1198
PPCLR-L	0.3891	0.1714	0.1198
PPCLR-R	1.3471	0.2584	1.3229
PPCLR-S	0.3787	0.1703	0.1195

표 2: type 2에서 $p = 10$이고 n이 증가할 때 추정오차의 평균.

gam=2	$n = 100$	$n = 500$	$n = 1000$
PLSLR	0.5127	0.3526	0.2847
PLR-A	0.3736	0.1564	0.1080
PLR-L	0.3743	0.1564	0.1079
PLR-R	1.2407	0.2824	1.2143
PLR-S	0.3691	0.1541	-
PPCLR-A	0.3745	0.1566	0.1078
PPCLR-L	0.3755	0.1572	0.1079
PPCLR-R	1.2411	0.3088	1.2216
PPCLR-S	0.3684	0.1545	0.1069

표 3: type 3에서 $p = 10$이고 n이 증가할 때 추정오차의 평균.

gam=3	$n = 100$	$n = 500$	$n = 1000$
PLSLR	0.4676	1.8961	0.2468
PLR-A	0.3781	1.9233	0.1130
PLR-L	0.3800	1.9147	0.1131
PLR-R	1.2848	1.7937	1.2605
PLR-S	0.3708	1.9226	0.1123
PPCLR-A	0.3762	1.9174	0.1122
PPCLR-L	0.3784	1.9125	0.1119
PPCLR-R	1.2860	1.7669	1.2680
PPCLR-S	0.3665	1.9240	-

표 4: type 1에서 $n = 100$이고 p이 증가할 때 추정오차의 평균.

gam=1	$p = 50$	$p = 100$	$p = 150$
PLSLR	0.8246	0.8448	0.8716
PLR-A	0.8663	2.4229	2.9370
PLR-L	0.8702	2.4882	2.9458
PLR-R	3.0123	1.4397	1.6629
PLR-S	0.8467	2.3932	-
PPCLR-A	0.8033	0.5700	0.5697
PPCLR-L	0.8062	0.6463	0.5284
PPCLR-R	3.0130	1.4384	1.7762
PPCLR-S	0.7874	0.7742	-

표 5: type 2에서 $n = 100$이고 p이 증가할 때 추정오차의 평균.

gam=2	$p = 50$	$p = 100$	$p = 150$
PLSLR	1.1465	0.9348	0.9543
PLR-A	0.8353	2.3151	2.7752
PLR-L	0.8370	2.3685	2.8551
PLR-R	2.7744	1.3801	1.6655
PLR-S	0.8253	-	-
PPCLR-A	0.8373	0.7948	0.8182
PPCLR-L	0.8397	0.8896	0.8121
PPCLR-R	2.7751	1.3795	1.7625
PPCLR-S	0.8237	-	-

표 6: type 3에서 $n = 100$이고 p이 증가할 때 추정오차의 평균.

gam=3	$p = 50$	$p = 100$	$p = 150$
PLSLR	1.0455	0.9025	0.9225
PLR-A	0.8456	2.3681	2.8394
PLR-L	1.3539	1.9517	2.3073
PLR-R	2.8729	1.4030	1.6708
PLR-S	0.8291	-	-
PPCLR-A	0.8412	0.7732	0.7778
PPCLR-L	0.8461	0.7870	0.6975
PPCLR-R	2.8756	1.4023	1.7734
PPCLR-S	0.8194	-	-