Journal of the Korean Data & Information Science Society
2014, **25**(1), 19–26

http://dx.doi.org/10.7465/jkdi.2014.25.1.19
한국데이터정보과학회지

연속형의 텐서곱과 범주형의 직합을 사용한 다항 로지스틱 회귀모형

심송용[1] · 강희모[2]

[12]한림대학교 금융정보통계학과

접수 2013년 9월 30일, 수정 2013년 11월 4일, 게재확정 2013년 11월 11일

요 약

다항 로지스틱 회귀모형의 설명변수가 연속형과 범주형을 모두 포함할 때 범주형 설명변수는 직합을 적용하고 연속형 설명변수는 텐서곱을 적용하는 모형을 제안한다. 변수선택의 기준으로 BIC를 사용하고, 제안된 모형의 알고리즘을 구현하였다. 구현된 알고리즘을 실제 자료에 적용하여 기존의 방법과 비교하여 제안된 모형이 더 좋은 분류율을 보임을 확인하였다.

주요용어: 검정자료, 정분류율, 훈련자료.

1. 서론

회귀모형에서 설명력 있는 변수를 선택하는 것은 주어진 변수만으로 모형의 성능을 향상시킬 수 있는 방법 중의 하나로 변수선택에 대한 연구가 다양하게 이루어지고 있다 (Kahng과 Shin, 2012; Choi와 Park, 2013; Shim과 Seok, 2013).

종속변수가 범주형일 때 적용할 수 있는 일반화선형모형의 연구는 Agarwal과 Studden (1980), Koo와 Lee (1994), Stone (1994) 등이 있으며 McCullagh와 Nelder (1989)는 많은 모형을 정리하였다.

일반화선형모형에서 공변량의 수가 커지면 다차원의 저주 (curse of dimensionality)를 피할 수 없으므로 Hastie와 Tibshirani (1990)는 주효과만 고려하여 일반화 가법모형 (generalized additive models: GAMs)을 제안하였다. Friedman (1991)은 다차원의 저주를 피하면서 주효과만 고려하는 GAMs의 한계를 극복하는 방법으로 MARS (multivariate adaptive regression spline)를 제안하였다.

이항 로지스틱 회귀모형에서 회귀계수 추정과 변수선택 과정 중 모형의 성능을 향상시키는 방법 연구되고 있으며 (Choi와 Park, 2013; Kahng 등, 2010; Kahng, 2011; Shim과 Seok, 2012), 다차원의 저주를 피하면서 여러 공변량에 텐서곱을 추가한 로지스틱 회귀모형도 소개되었다 (Lee 등, 2004). 또한 다항 자료인 경우도 Kooperberg 등 (1997)이 텐서곱을 사용한 모형의 계산방법을 구현하였다. Arppe (2012), Kooperberg (2013)에서 이와 관련된 패키지를 제공하고 있으며 연속형 자료 또는 연속형 자료의 텐서곱을 가법모형에 추가할 수 있도록 하였다.

이 논문에서는 일반화선형모형의 한 종류로 종속변수 Y가 다항 자료인 다항 로지스틱 회귀모형에서 주어진 독립변수가 범주형일 경우 직합 (direct sum)을 연속형인 경우 텐서곱 (tensor product)을 독립

[1] (200-702) 강원도 춘천시 한림대학길 1번지, 한림대학교 금융정보통계학과, 교수.
[2] 교신저자: (200-702) 강원도 춘천시 한림대학길 1번지, 한림대학교 금융정보통계학과, 겸임교수.
E-mail: hmkang@hallym.ac.kr

변수에 추가하고, 이 경우 과도하게 많아지는 독립변수의 갯수를 조절하기 위해 변수선택 기법을 제안한다.

제안한 방법을 실제 자료에 적용하여 다차원의 저주를 극복하고 판별율을 높이는 것을 확인하였다. 기존 모형에서 판별율은 모형을 적합한 자료로 계산하기 때문에 과적합이 발생할 수 있으므로 제안한 모형에서는 기존 모형에서 판별율을 구하는 방법과 데이터 마이닝 기법에 사용하는 방법인 훈련자료 (training data set)로 모형을 구하고 검정자료 (test data set)로 판별율을 계산하는 기법을 구현하였다.

2. 다항 로지스틱 회귀모형

K개의 수준을 갖고 있는 명목형 반응변수 Y와 이를 설명하는 M개의 설명변수 $x_1, x_2, \ldots, x_M$을 갖는 다항 로지스틱 회귀모형을 고려하자. 반응변수 Y가 택하는 값의 집합은 $\mathcal{K} = \{1, \ldots, K\}$로 나타낼 수 있으며 설명변수 $\mathbf{x} = (x_1, \ldots, x_M)$가 택하는 값의 집합은 $\mathbb{R}^M$의 부분집합인 $\mathcal{X}$로 나타내기로 한다. 이때 설명변수의 확률변수를 $\mathbf{X}$라고 하면 반응변수와 설명변수는 확률변수 쌍 $(\mathbf{X}, Y)$를 구성한다. $x \in \mathcal{X}$ 이고 $k \in \mathcal{K}$에 대하여 조건부 확률 $P(Y = K|\mathbf{X} = \mathbf{x})$가 양수 값을 가진다고 가정할 때

$$\theta(k|\mathbf{x}) = \log \frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K|\mathbf{X} = \mathbf{x})}, \quad \mathbf{x} \in \mathcal{X}, \quad k \in \mathcal{K} \tag{2.1}$$

라 하고 하면 $\theta(K|\mathbf{x}) = 0$ 이다. 식 (2.1)에 지수를 취하면

$$\frac{P(Y = k|\mathbf{X} = \mathbf{x})}{P(Y = K|\mathbf{X} = \mathbf{x})} = \exp \theta(k|\mathbf{x})$$

이므로, 양변을 각각 k에 대하여 합하면 $\sum_k P(Y = k|\mathbf{X} = \mathbf{x}) = 1$이므로

$$P(Y = K|\mathbf{X} = \mathbf{x}) = \frac{1}{\sum \exp \theta(k|\mathbf{x})}$$

이다. 따라서 주어진 $\mathbf{x}$에서 Y의 조건부 확률은

$$P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\exp \theta(k|\mathbf{x})}{\exp \theta(1|\mathbf{x}) + \cdots + \exp \theta(K|\mathbf{x})}, \quad \mathbf{x} \in \mathcal{X}, \quad k \in \mathcal{K} \tag{2.2}$$

로 나타낼 수 있다. 식 (2.2)를 다항 로지스틱 회귀모형 (polychotomous regression model, multinomial logistic regression model)이라 하고, 특별히 $K = 2$인 경우를 로지스틱 회귀모형 (logistic regression model)이라 부른다 (Kooperberg 등, 1997). $\theta(k|\mathbf{X} = \mathbf{x})$는

$$\theta(k|\mathbf{X} = \mathbf{x}) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kM}x_M, \quad 1 \le k \le K$$

로 선형가법모형이 된다. 또한 추정하려는 모형 $\theta(k|\mathbf{X} = \mathbf{x})$에 연속형 변수의 텐서곱과 범주형 변수의 직합을 고려한다면 다음과 같이

$$\begin{aligned} \theta(k|\mathbf{X} = \mathbf{x}) \quad = \quad & \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kM}x_M + \\ & \beta_{k12}x_1x_2 + \cdots + \beta_{k(M-1)M}x_{M-1}x_M, \quad 1 \le k \le K \end{aligned}$$

로 표현할 수 있다. 여기서 텐서의 성분은 각 변수와 스플라인 (spline) 함수로 구성되었다. 스플라인 함수는

$$(x_i - t_{ib})_+ = \begin{cases} x - t_b, & x_i \ge t_b \\ 0, & \text{그 외의 경우} \end{cases}$$

로 정의하는데 t_b들을 절단점 (break point) 또는 매듭점 (knot point)이라 한다 (Koo와 Lee, 1994). 매듭점 t_b의 선택은 $b = 1, 2, \cdots, n$과 $\alpha = 0.05$ 및

$$m = \left[\frac{-\log\left(-\frac{\log(1-\alpha)}{n}\right)}{\log(2) \times 5} \right] \tag{2.3}$$

인 (Friedman과 Silverman, 1989) b, α, m에 대해서 $1 + bm$번째 x들 중 BIC (Bayesian information criterion)를 기준으로 한 자료 선택 단계를 거쳐 최종 매듭점을 선택하여 다중공선성이 발생하지 않도록 하였다. 또한 변수선택 기준은 BIC를 적용하였다. BIC는 p개의 설명변수가 있을 때

$$\text{BIC} = -2l(\boldsymbol{\beta}) + p \cdot \ln(n) \tag{2.4}$$

이다 (Priestley, 1981, p. 375-376). 여기서 $l(\boldsymbol{\beta})$는 로그가능도 (log-likelihood)로

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\theta(Y_i|\mathbf{X}_i;\boldsymbol{\beta}) - c(\mathbf{X}_i;\boldsymbol{\beta})]$$

이며, $c(\mathbf{X}_i;\boldsymbol{\beta})$는

$$c(\mathbf{X}_i;\boldsymbol{\beta}) = \log[\exp\theta(1|\mathbf{X}_i;\boldsymbol{\beta}) + \cdots + \exp\theta(K|\mathbf{X}_i;\boldsymbol{\beta})]$$

이고, n은 표본수이다.

2.1. 범주형 자료의 직합 변환

회귀분석에서 범주형 변수를 설명변수로 사용할 경우 원래 변수의 값을 그대로 사용할 수 없는 경우가 많다. 이런 경우 원래의 변수를 여러개의 가변수 (dummy variable)로 변환하여 생성된 가변수를 설명변수로 사용한다.

범주형 변수가 갖는 값이 $1, 2, \ldots, C$중의 하나로써 범주의 수가 C개인 경우 가변수는 $(C-1)$개가 만들어지며 이 때 각 가변수는 $D_c = I_{[x_i=c]}$로 1 또는 0의 값을 갖는다 ($c = 1, 2, \ldots, C-1$).

이 논문에서는 범주형 자료를 모형에 추가할 때 앞의 일반적인 가변수 방법 대신에 직합을 사용하여 각 범주에 대한 가변수와 두개 이상의 범주가 합해진 가변수를 모형에 추가한다. 직합을 사용한 가변수의 생성은 다음과 같다.

1. 범주의 갯수가 C이면 가변수는 $2^C - 2$개를 만든다.
2. 범주형 자료를 가변수로 변경하기 위하여 1에서 $2^C - 2$ 까지 자연수를 모두 이진수로 변환한다. 이진수의 최대 자릿수는 2^{C-1}이며 이진수의 숫자 갯수는 이진수 2^0 자리에서 이진수 2^{C-1} 자리까지 C 개이다.
3. 십진수 A를 변환한 이진수가 $b_{C-1}2^{C-1} + \cdots + b_{c-1}2^{c-1} + \cdots + b_0 2^0$일 때 b_{c-1}은 가변수에서 범주 c에 대하여 1 또는 0을 결정한다 ($0 \le c \le C-1$).
4. A 번째 가변수 D_A는 십진수 A를 이진수 $\sum_{i=0}^{C-1} I_{[b_i=1]}2^i$로 표현할 때 십진수인 범주형 자료는 가변수 $D_A = \begin{cases} 1, & I_{[b_i \in 1]} \ i = 0, \cdots, C-1 \\ 0, & \text{그 외} \end{cases}$ 로 변환한다. 즉 범주형 자료 c는 가변수의 이진수 b_{c-1}로 결정되며 $b_i = 1, \ i \in (0, C-1)$인 범주는 가변수 변환 값이 모두 1 이고, 그 이외는 0 이다.

   예를 들어 범주가 C개인 경우

Table 2.1 Construction of dummy variable with C categories

Category C	dummy D_1	$\cdots$	dummy D_3	$\cdots$	dummy D_{2^C-2}
1 $\vdots$ c $\vdots$ C	$1_{(2)}, I_{[c=1]}$ $\begin{cases} 1, & c=1 \\ 0, & c\neq 1 \end{cases}$	$\cdots$	$11_{(2)}, I_{[c=1,2]}$ $\begin{cases} 1, & c=1,2 \\ 0, & \text{o.w} \end{cases}$	$\cdots$	$11\overbrace{\cdots}^{1}0_{(2)}, I_{[c\neq 1]}$ $\begin{cases} 1, & c\neq 1 \\ 0, & c=1 \end{cases}$

Table 2.1에서 가변수 D_1은 $b_0=1$인 경우로, 범주가 $c=1$인 경우만 가변수 값이 1 이고 그 이외의 경우는 0 이며, 가변수 D_3은 $b_0=1$, $b_1=1$인 범주 $c=1$과 범주 $c=2$일 때 가변수 값이 1 이고 그 이외의 경우는 0 이며 맨 끝에 만드는 2^C-2 번째 가변수 D_{2^C-2}는 이진수로 변환할 경우 $11\overbrace{\cdots}^{1}0_{(2)}$이므로 $c=1$인 경우만 가변수 값이 0 이고 나머지 범주값 $c\neq 1$은 가변수 값이 1 이다. 예를 들어 범주의 갯수가 $C=3$일 때 가변수로 변환하면

1. 가변수의 갯수는 $2^3-2=6$개이다.
2. 범주 $C=3$에 대한 가변수의 이진수 범위는 $1\ (001_{(2)}) \sim 2^3-2=6\ (110_{(2)})$ 이다.
3. 범주형 자료 c는 다음과 같이 가변수 조건에 맞게

$$
\begin{aligned}
D_1 = 1(001_{(2)}) = I_{[c=1]} &= \begin{cases} 1, & c=1 \\ 0, & \text{그 외} \end{cases} \\
D_2 = 2(010_{(2)}) = I_{[c=2]} &= \begin{cases} 1, & c=2 \\ 0, & \text{그 외} \end{cases} \\
D_3 = 3(011_{(2)}) = I_{[c=1,2]} &= \begin{cases} 1, & c=1,2 \\ 0, & \text{그 외} \end{cases} \\
D_4 = 4(100_{(2)}) = I_{[c=3]} &= \begin{cases} 1, & c=3 \\ 0, & \text{그 외} \end{cases} \\
D_5 = 5(101_{(2)}) = I_{[c=1,3]} &= \begin{cases} 1, & c=1,3 \\ 0, & \text{그 외} \end{cases} \\
D_6 = 6(110_{(2)}) = I_{[c=2,3]} &= \begin{cases} 1, & c=2,3 \\ 0, & \text{그 외} \end{cases}
\end{aligned}
$$

로 값을 설정한다.

범주가 3개인 경우 직합을 고려한 가변수 변환은 다음과 같다.

$$
\begin{array}{c} \text{원자료} \\ \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \end{pmatrix} \end{array}
\rightarrow
\begin{array}{c} \begin{matrix} D_1 & D_2 & D_3 & D_4 & D_5 & D_6 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}
$$

2.2. 연속형 설명변수의 텐서곱과 범주형 설명변수의 직합 적용

모형에 포함될 설명변수가 연속형일 경우 변수값의 텐서곱을 범주형인 경우 직합을 얻어 이를 모형에 포함시킨다. 연속형 변수의 텐서곱과 범주형 변수의 직합은 다음과 같은 순서로 얻어진다.

1. 연속형 자료는 가장 먼저 x_i $(i = 1, ..., p)$에서 모형에 기저 (basis)를 추가한다.
2. 범주형 자료는 범주 C에 대하여 직합으로 $2^C - 2$개의 가변수를 만들어 모형에 추가한다.
3. 연속형 변수에서 이미 모형에 추가된 x_i 중 $(x_i - t_{ib})_+$를 추가한다.
4. 연속형 변수에서 이미 모형에 추가된 x_i와 x_j는 텐서곱 $x_i x_j$를 추가한다.
5. 연속형 변수에서 이미 모형에 추가된 x_i와 $(x_i - t_{ib})_+$는 텐서곱 $x_i(x_i - t_{ib})_+$를 추가한다.
6. 연속형 변수에서 이미 모형에 추가된 $(x_i - t_{ib})_+$와 $(x_j - t_{jb})_+$는 텐서곱 $(x_i - t_{ib})_+(x_j - t_{jb})_+$를 추가한다.

모형에 변수선택은 식 2.4의 BIC 값이 감소하다가 증가하면 중지한다.

3. 제안된 모형의 구현 및 성능평가 적용 사례

앞에서 설명된 방법으로 범주형 설명변수의 직합 및 연속형 설명변수 텐서곱을 사용한 모형과 기존 모형의 비교를 위하여 실제자료를 두 모형에 적용하여 정분류율을 계산하였다.

사용한 자료는 모 통신회사의 자료로 세개의 고객등급을 종속변수로, 독립변수는 2개의 연속형 변수와 4개의 범주형 변수로 구성되었으며 표본크기는 각 고객등급당 각각 9,735개, 10,510개, 9,755개로 전체 30,000개이다.

본 논문에서 제안한 모형의 성능을 평가하기 위하여 SPSS에서 제공하는 기존 다항 로지스틱 회귀모형을 적용한 결과를 비교하였다.

제안된 모형의 알고리즘은 C 언어로 구현하였고 컴파일러는 gcc 4.6.2를 사용하였으며 SPSS는 64bit형 Version 20을 사용하였다.

3.1. 추정 모형 및 모형 성능

제안된 다항 로지스틱회귀모형의 경우 추가된 기저는 연속형이 2개의 텐서, 20개의 텐서 스플라인, 7개의 텐서곱 스플라인 그리고 범주형에 대한 6개의 직합까지 총 35개가 제안된 알고리즘에 의해서 선택되었다.

제안된 모형에서 사용한 2개의 연속형 변수와 4개의 범주형 변수 설명변수를 모두 사용하여 모형에 포함하고, 기존의 모형과 비교하고자 한다. SPSS의 결과 2개의 연속형 변수와 4개의 범주형 변수가 모두 유의하였다.

기존 모형과 제안된 모형의 성능을 평가하기 위하여 각 모형에서의 관측값 예측값에 따른 정분류율 (correct classification rate)을 살펴보기로 하자. SPSS로 추정한 다항 로지스틱 회귀모형은 정분류율을 계산한 결과 59.6%이고 (Table 3.1) 제안한 모형의 정분류율은 66.0% (Table 3.2)로 제안한 모형의 판별이 6.4% 포인트 더 높게 나타났다. 또한 등급별 정분류율은 기존 모형의 경우 73.1%, 45.8%, 61.1%로 두 번째 등급이 다른 등급보다 저조한 판별 결과를 보였고, 제안된 모형의 경우 71.1%, 62.6%, 64.6% 등으로 각 등급마다 기존 모형보다 고른 판별 결과가 나타났다.

Table 3.1 Classification of usual model

observed \ predicted	1	2	3	correct classification
1	7114	2200	421	73.1%
2	2825	4810	2875	45.8%
3	168	2112	5958	61.1%
total (%)	38.7%	30.4%	30.8%	59.6%

Table 3.2 Classification of proposed model

observed \ predicted	1	2	3	correct classification
1	6921	2230	584	71.1%
2	1495	6580	2435	62.6%
3	454	2997	6304	64.6%
total (%)	29.6%	39.4%	31.1%	66.0%

3.2. 데이터 마이닝 기법으로 모형 추정 및 모형 성능

자료의 일부로 모형을 설정하고, 나머지의 자료로 모형의 성능을 평가하는 데이터 마이닝 방법으로 제안된 모형의 성능을 알아보기로 한다. 전체 자료에서 80%는 훈련자료로 모형 구축에 사용하고 나머지 20%는 검정자료로 모형 적합도를 계산하였다.

훈련자료로 모형을 적합한 결과 연속형이 2개의 텐서, 23개의 텐서 스플라인, 9개의 텐서곱 스플라인 그리고 범주형에 대한 6개의 직합까지 총 40개가 알고리즘에 의해서 선택되었다. 그리고 검정자료로 정분류율을 구한 결과 66.1% (Table 3.3)이며 훈련자료와 검정자료로 모형 구축과 정분류율을 계산한 것이 전체자료로 모형과 정분류율을 구한 경우 보다 약 0.1% 크게 나타났다.

연속형 변수의 텐서를 모형에 추가할 때 기저 후보는 식 2.3을 기준으로, 범주형 자료는 직합으로 모형의 변수선택을 결정하기 때문에 훈련자료와 검정자료의 랜덤화에 따라 모형에 추가되는 기저가 다를 수 있다. 그렇기 때문에 검정자료에 대한 정분류율은 전체자료로 구한 정분류율보다 크거나 또는 작은 값이 나올 수 있다. 실제 이와 같은 훈련자료와 검정자료를 여러번 반복하여 제안된 알고리즘을 적합하여도 Table 3.3과 비슷한 결과가 관찰되었으며 제안한 모형보다 추가된 기저의 갯수가 많거나 정분류율이 높은 경우도 있었다.

Table 3.3 Classification of proposed model with a data mining method

observed \ predicted	1	2	3	correct classification
1	1395	487	139	69.0%
2	269	1353	474	64.6%
3	85	615	1280	64.6%
total (%)	28.7%	40.3%	31.0%	66.1%

4. 맺음말

다항 로지스틱 회귀분석은 사후확률을 직접 모형화하기 때문에 다항 범주형 자료의 판별방법으로 사용되고 있다. 본 논문에서 구현한 모형은 텐서 스플라인과 직합을 추가할 수 있어서 인접한 값이 텐서에 계속 추가되면 과적합이 발생할 수 있으므로 이를 피할 수 있는 변수선택 기법을 추가하였다. 본 연구에는 주어진 설명변수 내에서 새로운 설명변수를 추가 생성하므로 주어진 정보를 최대로 활용한다는 이점이 있으며 실자료에 적용하여 정분류율이 높아짐을 확인하였다.

References

Agarwal, G. G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing spline. *The Annals of Statistics*, **8**, 1307-1325.

Arppe, A. (2012). polytomous: Polytomous logistic regression for fixed and mixed effects. R package version 0.1.4., `http://CRAN.R-project.org/package=polytomous`.

Choi, S. and Park, C. (2013). An educational tool for regression models with dummy variables using Excel VBA. *Journal of the Korean Data & Information Science Society*, **24**, 593-601.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, **19**, 1-141.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*, **31**, 3-39.

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

Kahng, M. (2011). A study on log-density ratio in logistic regression model for binary data. *Journal of the Korean Data & Information Science Society*, **22**, 107-113.

Kahng, M. and Shin, E. (2012). A study on log-density with log-odds graph for variable selection in logistic regression. *Journal of the Korean Data & Information Science Society*, **23**, 99-111.

Kahng, M., Kim, B. and Hong, J. (2010). Graphical regression and model assessment in logistic model. *Journal of the Korean Data & Information Science Society*, **21**, 21-32.

Koo, J. and Lee, Y. (1994). Bivariate B-splines in generalized linear models. *Journal of Statistical Computation and Simulation*, **50**, 119-129.

Kooperberg, C. (2013). polspline: Polynomial spline routines. R package version 1.1.8., `http://CRAN.R-project.org/package=polspline`.

Kooperberg, C., Bose, S. and Stone, J. (1997). Polychotomous regression. *Journal of the American Statistical Association*, **92**, 117-127.

Lee, S., Sim, S. and Koo, J. (2004). A study on data mining using the spline basis. *Communications of the Korean Statistical Society*, **11**, 255-264.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2nd ed., Chapman and Hall, London.

Priestley, M. B. (1981). *Spectral analysis and time series*, Academic Press, London.

Shim, J. and Seok, K. (2012). Semiparametric kernel logistic regression with longitudinal data. *Journal of the Korean Data & Information Science Society*, **23**, 385-392.

Shim, J. and Seok, K. (2013). GACV for partially linear support vector regression. *Journal of the Korean Data & Information Science Society*, **24**, 391-399.

Stone, C. J. (1994). The use of polynomial splines and their products in multivariate function estimation. *The Annals of Statistics*, **22**, 118-171.

Journal of the Korean Data & Information Science Society
2014, **25**(1), 19–26

http://dx.doi.org/10.7465/jkdi.2014.25.1.19
한국데이터정보과학회지

A polychotomous regression model with tensor product splines and direct sums

Songyong Sim[1] · Heemo Kang[2]

[12]Department of Finance & Information Statistics, Hallym University

Received 30 September 2013, revised 4 November 2013, accepted 11 November 2013

Abstract

In this paper, we propose a polychotomous regression model when independent variables include both categorical and numerical variables. For categorical independent variables, we use direct sums, and tensor product splines are used for continuous independent variables. We use BIC for varible selections criterior. We implemented the algorithm and apply the algorithm to real data. The use of direct sums and tensor products outperformed the usual multinomial logistic regression model.

Keywords: BIC, classification rate, test data, training data.

[1] Professor, Department of Finance & Information Statistics, Hallym University, Chuncheon 200-702, Korea.
[2] Corresponding author: Adjunct professor, Department of Finance & Information Statistics, Hallym University, Chuncheon 200-702, Korea. E-mail: hmkang@hallym.ac.kr