응용통계연구 제 **14**권 *1*호, *2001*년, *pp. 71–80*

로지스틱 회귀모형을 이용한 비대칭 종형 확률곡선의 추정*

박성현[1] 김기호[2] 이소형[3]

요 약

로지스틱 회귀모형은 이항 반응자료에 대한 가장 보편적인 일반화 선형모형으로 독립변수에 대한 확률함수를 추정하는데 이용된다. 많은 실제적 상황에서 확률함수가 종형의 곡선형태로 표현되는데 이 경우에는 2차항을 포함한 로지스틱 회귀모형을 이용하여 확률함수를 추정할 수 있다. 그러나 2차항을 포함한 로지스틱 회귀모형을 이용한 분석은 대칭성을 갖는 확률함수에 대한 가정으로 인해 비대칭 형태의 종형곡선에서는 확률함수의 신뢰성이 저하되고, 2차항을 포함하기 때문에 독립변수의 효과를 설명하기가 쉽지 않다는 제한점을 가지고 있다. 본 논문에서는 이러한 문제점을 해소하기 위해서 로지스틱 회귀분석과 반복적 이분법을 이용하여 종형의 형태에 관계없이 확률곡선을 추정하는 방법론을 제안하고 모의 실험을 통해 2차항을 포함한 로지스틱 회귀모형과 비교하고자 한다.

주요용어: 로지스틱 회귀모형, 비대칭 종형, 확률곡선 추정.

1. 서론

독립변수들로부터 이항 반응자료에 대한 확률함수를 추정하는데 이용되는 로지스틱 회귀모형은 Cox(1970)가 로지스틱 회귀분석방법을 처음 제시한 이후로 Agreti(1990)는 로지스틱 회귀모형을 가장 보편적인 일반화 선형모형이라 하였으며, 의학이나 생물학 등 여러 분야에 널리 이용되고있다. 요즈음에 와서는 Berry와 Linoff(1997)의 대용량 데이터에근거하는 데이터마이닝 기법(data mining technique)들의 발달로 기업이나 개인의 신용평가, 리스크 분석, 고객 점수(scoring)산정 등에 있어 로지스틱 회귀모형이 많이 응용되고 있다. 일반적으로 지수에 1차항만을 포함하는 1차 로지스틱 회귀모형은 확률함수에 대해 단조증가나 단조감소가 나타나는 S-곡선의 형태로 나타내어진다. 하지만 많은 실제적 상황에서 확률함수가 종형의 곡선형태로 표현되는데 이 경우에는 다음과 같이 지수에 2차항을 포함한 2차 로지스틱 회귀모형을 이용하여 확률함수를 추정할 수 있다.

$$logit(\pi(x)) = \alpha + \beta_1 x + \beta_2 x^2$$

* 이 논문은 1999년도 두뇌한국 21사업 핵심분야에 의하여 지원되었음.

1) (151-742) 서울시 관악구 신림동, 서울대학교 통계학과, 교수
E-mail: parksh@plaza.snu.ac.kr
2) (135-798) 서울시 강남구 삼성동 159-1, PWC Consulting, 수석컨설턴트
E-mail: kiho.kim@kr.pwcglobal.com
3) (136-716) 서울시 서초구 방배동 935-34, 외환카드 IT팀, 대리

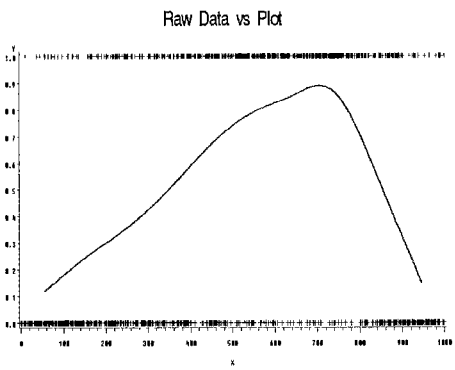


그림 1.1: 원자료 vs 산점도

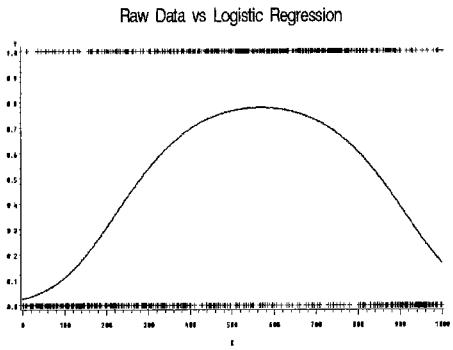


그림 1.2: 2차 로지스틱 회귀

그러나 이러한 2차 로지스틱 회귀모형을 이용한 분석은 다음과 같은 제한점이 있다. 첫째는 이론적으로 2차 로지스틱 회귀모형은 대칭성을 갖는 확률함수를 가정한다는 것이다. 즉, 종형곡선의 형태가 대칭인 경우에 대해서는 확률함수를 잘 추정하지만 비대칭 형태의 종형곡선에서는 추정된 확률함수의 신뢰성이 저하된다는 것이다. 둘째는 추정된 확률함수가 독립변수의 2차항을 포함하므로 독립변수의 효과를 설명하기가 쉽지 않다는 것이다. 이와 같이 확률분포의 형태가 비대칭 형태의 종형인 경우에 대해서 로지스틱 회귀모형이 어떻게 적합되는지 살펴보기 위해서 확률 분포의 형태가 비대칭 형태가 되도록 SAS 프로그램을 이용해 임의로 추출된 1000개의 자료를 이용하였다.

그림 1.1에서는 확률분포의 형태가 비대칭 종형인 경우로 추출된 원자료(raw data)와 자료의 전체적인 개형과 적합한 로지스틱 회귀모형을 파악하기 위해 Hosmer와 Lemeshow (1989)가 제안한 확률 산점도를 함께 나타낸 그림이다. 원자료를 살펴보면 반응변수에 대해 가운데 부분에선 1값이 양쪽 끝 부분에선 0값이 많이 나타나고 있다. 확률 산점도는 원자료를 10개의 그룹으로 나누어 각각의 평균 확률값을 구하고 SAS의 gplot option중 splines를 사용하여 표현한 그림으로 확률곡선의 형태가 비대칭 종형임을 확인할 수 있다. 그림 1.2는 이 자료에 2차 로지스틱 회귀모형을 적합시킨 결과로써 그림 1.1의 확률 산점도와는 달리 종형형태를 나타내고는 있지만 비대칭 정도가 잘 나타나고 있지 않다. 위의 그림을 통해서 살펴보았듯이 2차 로지스틱 회귀모형의 경우 종형곡선의 형태가 비대칭 형태인 경우에 추정된 확률함수의 신뢰성이 저하됨을 알 수 있다. 본 논문에서는 이러한 문제점을 해소하기 위해서 1차 로지스틱 회귀분석과 반복적 이분법을 이용하여 종형의 형태에 관계없이 확률곡선을 추정하는 방법론을 제안하고 모의 실험을 통해 2차항을 포함한 로지스틱 회귀모형과 비교하고자 한다.

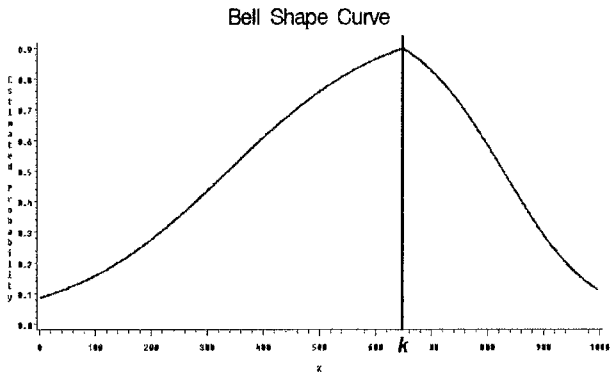


그림 2.1: 비대칭 종형 형태

2. 이분법을 이용한 종형 확률곡선의추정

2.1. 개요

확률의 분포형태가 종형인 경우를 살펴보자.

그림 2.1은 확률분포의 형태가 종형인 경우로 최대 확률값을 갖는 k 를 중심으로 k 보다 작은 구간에서는 단조증가가 k 보다 큰 구간에서는 단조감소의 형태가 나타남을 알 수 있다. 이러한 확률분포의 이론적 모형은 1차 로지스틱 회귀모형을 이용해 다음과 같이 표현될 수 있다.

$$\begin{aligned} \pi(x) &= \pi_1(x)^{I(x)}\pi_2(x)^{[1-I(x)]} \\ &= [\frac{\exp(\alpha_1+\beta_1 x)}{1+\exp(\alpha_1+\beta_1 x)}]^{I(x)}[\frac{\exp(\alpha_2+\beta_2 x)}{1+\exp(\alpha_2+\beta_2 x)}]^{[1-I(x)]} \end{aligned} \tag{2.1}$$

where $\pi_1(k)=\pi_2(k)$

$$I(x)=\begin{cases} 1, & \text{if} \quad x \le k \\ 0, & \text{if} \quad x > k \end{cases}$$

(식 2.1)의 전체 모형의 추정확률함수 $\hat{\pi}$ 은 $\widehat{\pi_1}(k^*)=\widehat{\pi_2}(k^*)$ 를 만족시키는 k^* 를 찾음으로써 구해질 수 있는데 전체 모형에 대한 회귀계수 추정량 $\widehat{\alpha_1}$, $\widehat{\beta_1}$, $\widehat{\alpha_2}$, $\widehat{\beta_2}$또한 k^* 에 의해 결정되므로 k^* 는 닫힌형식(closed form)으로 표현되지 않는다. 따라서 가장 적합한 k^* 의 값은 수치해석(Numerical Analysis)적인 방법을 이용해 찾을 수 있는데 본 연구에서는 Burden(1997)의 중간값 정리(Intermediate value theorem)에 기초한 방정식의 근을 찾는 방법인 이분법(bisection method)을 이용해 k^* 를 찾는 방법을 제안한다.

2.2. 이분법을 이용한 확률함수의 추정

이분법은 함수의 부호가 변화하는 구간을 찾아내어 근을 구하는 방법으로 이러한 이분

법을 이용해 k^* 를 구하는 과정은 다음과 같이 정리될 수 있다.

1단계. 임의의 초기값(x)에서 자료를 나누어 각각 1차 로지스틱 회귀모형을 적합시킨다. 이때 초기값으로 사용되는 x 값은 일반적으로 자료의 확률 산점도를 이용하여 가장 확률이 높은 값 을 대략적으로 구하여 이용할 수 있다.

2단계. 1차 로지스틱 회귀모형을 적합시켜서각각의 예측 확률값 $\widehat{\pi_1}(x), \widehat{\pi_2}(x)$을 구한다.

3단계. $\widehat{\pi_1}(x) > \widehat{\pi_2}(x)$이면 x 를 $\triangle$ 만큼 감소시킨다. (x')

$\widehat{\pi_1}(x) < \widehat{\pi_2}(x)$이면 x 를 $\triangle$ 만큼 감소시킨다.(x')

4단계. x'에서 다시 자료를 나누어 각각 1차 로지스틱 회귀모형을 적합시킨다.

5단계. x'에서의 예측 확률값 $\widehat{\pi_1}(x'), \widehat{\pi_2}(x')$ 을 구한다.

6단계. $\widehat{\pi_1}(x) > \widehat{\pi_2}(x)$ 이고 $\widehat{\pi_1}(x') > \widehat{\pi_2}(x')$ 이면 x' 를 $\triangle$ 만큼 감소시킨다.

$\widehat{\pi_1}(x) > \widehat{\pi_2}(x)$ 이고 $\widehat{\pi_1}(x') < \widehat{\pi_2}(x')$ 이면 x' 를 $\frac{\triangle}{2}$ 만큼 증가시킨다.

$\widehat{\pi_1}(x) < \widehat{\pi_2}(x)$ 이고 $\widehat{\pi_1}(x') < \widehat{\pi_2}(x')$ 이면 x' 를 $\triangle$ 만큼 감소시킨다.

$\widehat{\pi_1}(x) < \widehat{\pi_2}(x)$ 이고 $\widehat{\pi_1}(x') > \widehat{\pi_2}(x')$ 이면 x' 를 $\frac{\triangle}{2}$ 만큼 감소시킨다.

7단계. 위의 4단계~ 6단계를 반복하여 옮긴점에서의 예측 확률값과 이전의 확률값을 비교해 부호가 바뀌면 그 구간사이에서 $\triangle$를 $\frac{\triangle}{2}, \frac{\triangle}{4}, \frac{\triangle}{8}, \cdots$ 로 계속 반으로 감소시키면서 $\widehat{\pi_1}(x^*) - \widehat{\pi_2}(x^*)) \leq \epsilon$ 인 x^* 를 구한다. 이때 $\epsilon \geq 0$ 은 상수이고, 일반적으로 ϵ 값은 연구분야의 특성에 따라 적합하게 결정할 수 있다.

8단계. 위의 결과 얻어진 x^* 값을 k^* 값으로 한다.

위의 과정을 그림으로 표현하면 아래의 그림 2.2와 같이 나타낼 수 있으며, 여기서 $x = 630, \triangle = 120, k^* = 720$ 을 나타내고 있다.

앞에서 이분법을 이용하여 구한 k^* 값을 이용해 전체 확률함수를 추정하여 다음과 같이 표현할 수 있다.

$$\begin{aligned} \widehat{\pi}(x) &= \widehat{\pi_1}(x)^{I(x)}\widehat{\pi_2}(x)^{1-I(x)} \\ &= \left[\frac{\exp(\widehat{\alpha_1} + \widehat{\beta_1 x})}{1+\exp(\widehat{\alpha_1} + \widehat{\beta_1 x})}\right]^{I(x)} \left[\frac{\exp(\widehat{\alpha_2} + \widehat{\beta_2 x})}{1+\exp(\widehat{\alpha_2} + \widehat{\beta_2 x})}\right]^{1-I(x)} \end{aligned} \tag{2.2}$$

where $\widehat{\pi_1}(k^*) = \widehat{\pi_2}(k^*)$

$$I(x) = \begin{cases} 1, & \text{if} \quad x \leq k^* \\ 0, & \text{if} \quad x > k^* \end{cases}$$

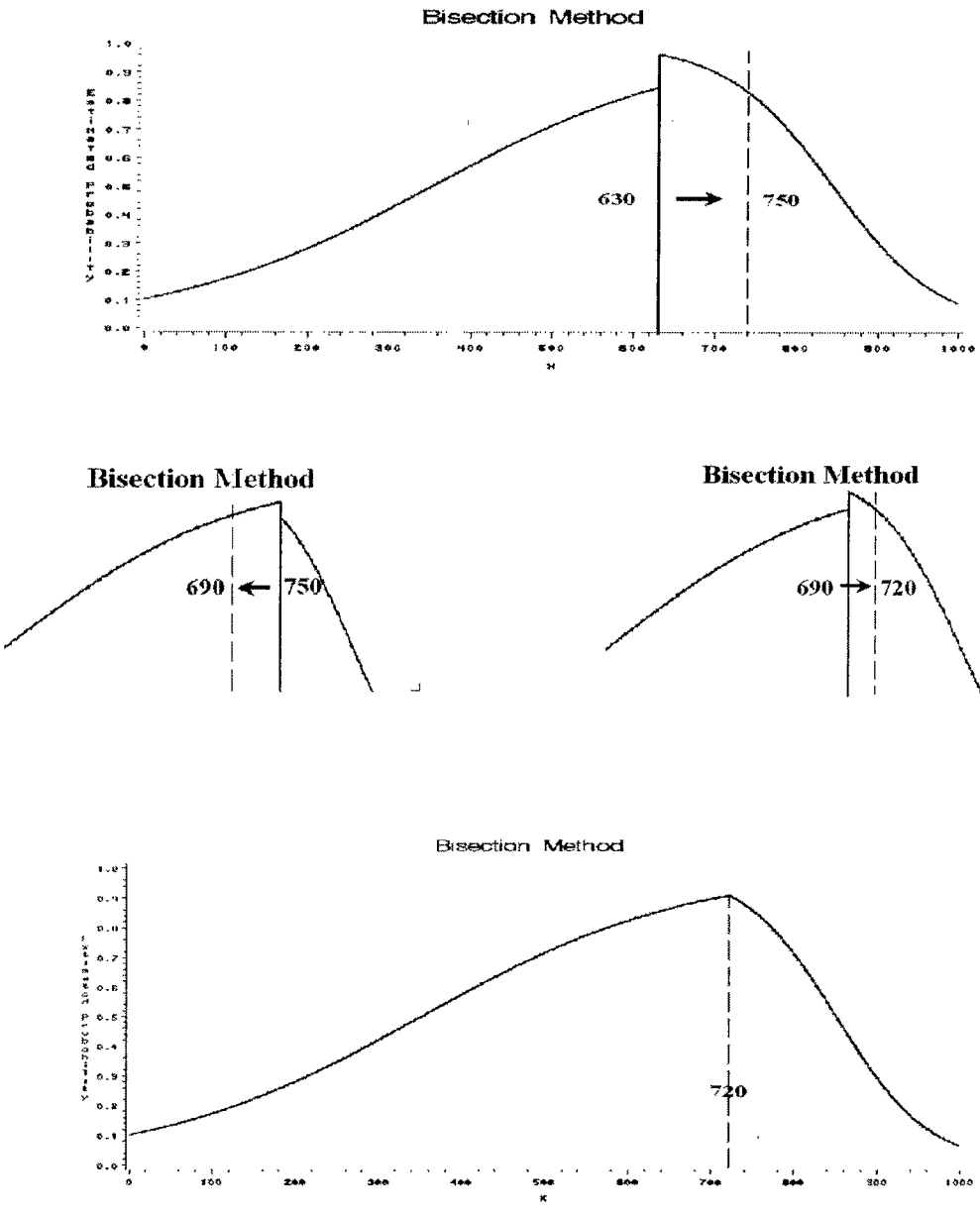

Bisection Method
630
750
Bisection Method
690
750
Bisection Method
690
720
Bisection Method
720

그림 2.2: 이분법

(식 2.2)에서 추정된 전체 확률함수의

$$H_0 : \begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} = 0, H_1 : \begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} \neq 0$$

에 대한 적합도는 우도비 검정(likelihood ratio test)을 이용한 G 통계량으로 다음과 같이 구할 수 있으며, (식 2.3)과 같이 주어진다.

$$\begin{aligned} G &= -2\text{Max}\ \ln L(\text{without the variable}) + 2\text{Max}\ \ln L(\text{with the variable}) \\ &= -2\ln L_0(\hat{\pi}(x)) + 2\ln L_1(\hat{\pi}(x)) \end{aligned}$$

$$\begin{aligned} L(\pi(x)) &= \prod_x (\pi(x)) \\ &= \prod_{x\le k^*} [\pi(x)] \prod_{x>k^*} [\pi(x)] \end{aligned}$$

$$L_1(\hat{\pi}(x)) = \prod_{x\le k^*} \widehat{\pi_1}(x)^y [1-\widehat{\pi_1}(x)]^{(1-y)} \prod_{x>k^*} \widehat{\pi_2}(x)^y [1-\widehat{\pi_2}(x)]^{(1-y)}$$

단, y는 0 또는 1 을 갖는 반응변수

$$L_0(\hat{\pi}(x)) = (\frac{n_1}{k^*})^{n_1}(1-\frac{n_1}{k^*})^{(k^*-n_1)} \cdot (\frac{n_2}{n-k^*})^{n_2}(1-\frac{n_2}{n-k^*})^{(n-k^*-n_2)}$$

단, $n_1 = \sum_{x\le k^*} y \qquad n_2 = \sum_{x>k^*} y$

$$\begin{aligned} G &= -2\ln(L_0/L_1) \\ &= -2\ln L_0 + 2\ln L_1 \\ &= -2\ln[(\frac{n_1}{k^*})^{n_1}(1-\frac{n_1}{k^*})^{(k^*-n_1)} \cdot (\frac{n_2}{n-k^*})^{n_2}(1-\frac{n_2}{n-k^*})^{(n-k^*-n_2)}] \\ &\quad +2\ln[\prod_{x\le k^*} \widehat{\pi_1}(x)^y [1-\widehat{\pi_1}(x)]^{(1-y)} \prod_{x>k^*} \widehat{\pi_2}(x)^y [1-\widehat{\pi_2}(x)]^{(1-y)}] \\ &= -2\ln[(\frac{n_1}{k^*})^{n_1}(1-\frac{n_1}{k^*})^{(k^*-n_1)}] - 2\ln[(\frac{n_2}{n-k^*})^{n_2}(1-\frac{n_2}{n-k^*})^{(n-k^*-n_2)}] \\ &\quad +2\ln[\prod_{x\le k^*} \widehat{\pi_1}(x)^y [1-\widehat{\pi_1}(x)]^{(1-y)}] + 2\ln[\prod_{x>k^*} \widehat{\pi_2}(x)^y [1-\widehat{\pi_2}(x)]^{(1-y)}] \\ &= -2\ln[(\frac{n_1}{k^*})^{n_1}(1-\frac{n_1}{k^*})^{(k^*-n_1)}] + 2\ln[\prod_{x\le k^*} \widehat{\pi_1}(x)^y [1-\widehat{\pi_1}(x)]^{(1-y)}] \\ &\quad -2\ln[(\frac{n_2}{n-k^*})^{n_2}(1-\frac{n_2}{n-k^*})^{(n-k^*-n_2)}] + 2\ln[\prod_{x>k^*} \widehat{\pi_2}(x)^y [1-\widehat{\pi_2}(x)]^{(1-y)}] \\ &= -2\ln[\frac{(\frac{n_1}{k^*})^{n_1}(1-\frac{n_1}{k^*})^{(k^*-n_1)}}{\prod_{x\le k^*} \widehat{\pi_1}(x)^y [1-\widehat{\pi_1}(x)]^{(1-y)}}] - 2\ln[\frac{(\frac{n_2}{n-k^*})^{n_2}(1-\frac{n_2}{n-k^*})^{(n-k^*-n_2)}}{\prod_{x>k^*} \widehat{\pi_2}(x)^y [1-\widehat{\pi_2}(x)]^{(1-y)}}] \end{aligned}$$

따라서

$$\begin{aligned} G &= 2\sum_{x\le k^*}[y\ln(\widehat{\pi_1}(x)) + (1-y)\ln(1-\widehat{\pi_1}(x))] \\ &\quad -2[n_1\ln(n_1) + (K^*-n_1)\ln(K^*-n_1) - k^*\ln k^*)] \\ &\quad +2\sum_{x>k^*}[y\ln(\widehat{\pi_2}(x)) + (1-y)\ln(1-\widehat{\pi_2}(x))] \\ &\quad -2[n_2\ln(n_2)(n-k^*-n_2)\ln(n-k^*-n_2) - (n-k^*)\ln(n-k^*)] \end{aligned} \tag{2.3}$$

(식 2.3)은 전체 모형의 적합도 통계량 G 로서 k^* 를 중심으로한 각각의 1차 로지스틱 회귀모형의 적합도 통계량의 합으로 표현됨을 알 수 있다. 따라서 각 1차 로지스틱 회귀모형의 적합도 통계량값을 $G(\widehat{\pi_1}), G(\widehat{\pi_2})$ 라 하면 전체 모형에 대한 적합도 통계량을 다음과 같이 나타낼 수 있다.

$$G(\hat{\pi}) = G(\widehat{\pi_1}) + G(\widehat{\pi_2}) \tag{2.4}$$

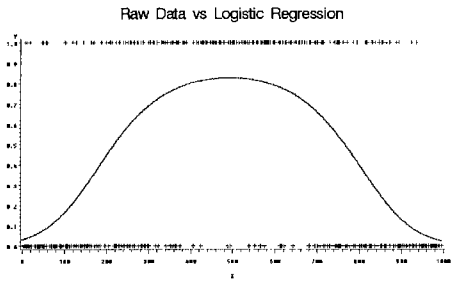
3. 모의실험

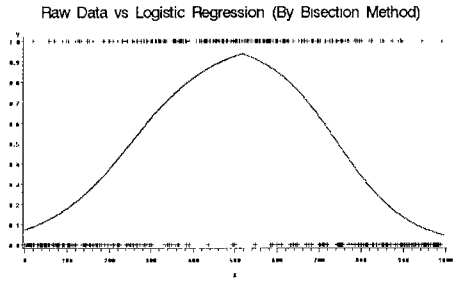
모의 실험에서는 확률곡선의 형태가 종형인 자료에 대해 앞에서 설명한 이분법(bisection method)과 2차 로지스틱 회귀모형을 적용하고 각각의 모형의 적합정도를 판단하기 위해 적합통계량인 G 값을 구해 비교해 보겠다.

1개의 설명변수 X 를 SAS 프로그램을 사용해 UNIFORM(0,1)로부터 n개를 임으로 추출하여 1000을 곱한 후 정수화하여 1부터 1000까지의 자료를 얻었으며, 반응변수는 Y는 B(1,p)로부터 각각 추출하였다. 여기서는 n을 500으로 하고 확률곡선이 종형인 경우에 대칭, 비대칭형태가 되도록 X 변수를 100단위로 구간을 나누어 p 값을 다르게 주었으며, k^* 값을 찾기위한 프로그램은 SAS Macro를 이용하였다.

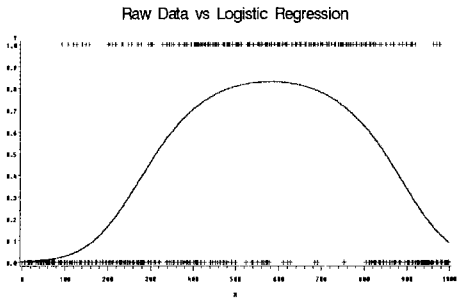
표 3.1: 2차 로지스틱과 이분법의 적합도 비교

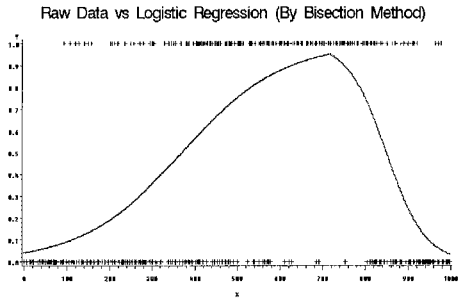
표본 크기	종형형태	G	
		2차로지스틱	이분법
n=500	(a) 대칭형태	192.163	206.478
	(b) 왼쪽으로 편중된 형태	208.462	233.850
	(c) 오른쪽으로 편중된 형태	184.122	221.183
n=300	(a) 대칭형태	156.794	172.931
	(b) 왼쪽으로 편중된 형태	157.640	181.749
	(c) 오른쪽으로 편중된 형태	133.077	162.547
n=100	(a) 대칭형태	71.888	74.828
	(b) 왼쪽으로 편중된 형태	81.204	86.341
	(c) 오른쪽으로 편중된 형태	100.306	162.547



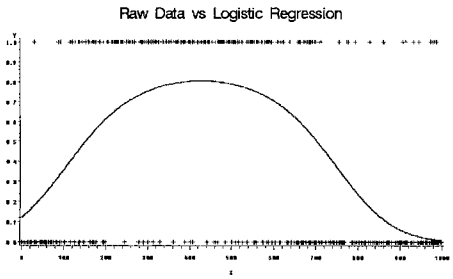


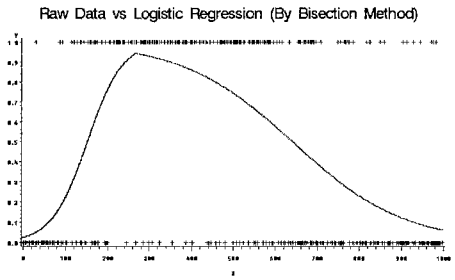
(a) 종형형태





(b) 오른쪽으로 편중된 형태





(c) 왼쪽으로 편중된 형태

그림 3.1: 각 형태별 2차 로지스틱 & 이분법의 확률곡선

그림 3.1과 표 3.1은 각 형태별 2차 로지스틱 확률곡선과 이분법 로지스틱 확률곡선에 대한 그래프와 G 값을 나타낸다. 그림 3.1에서 살펴보면 (a)대칭형태의 경우는 2차 로지스틱과 이분법을 이용한 확률곡선들의 형태가 크게 차이가 나지 않지만, (b)와 (c)의 경우와 같이 비대칭 자료의 경우에는 확률곡선들의 형태에 차이가 있음을 볼 수 있다.

표 3.1은 그림 3.1의 경우들에 대한 G 값을 정리한 것으로써 (a)의 대칭형태인 경우는 두 방법의 적합도가 크게 차이가 나지 않지만 (b)와 (c)의 비대칭 경우에는 2차 로지스틱과 비교해 이분법에 의한 방법의 적합도가 크게 높음을 알 수 있다. 또한 이러한 결과는 표본의 크기와 비대칭의 정도에 관계없이 비슷하게 나타남을 본 연구를 통해 알 수 있다.

4. 결론

위의 모의실험을 통해 종형의 각 형태별로 2차 로지스틱과 이분법을 적용해 보았다. 그 결과 이분법에 의한 로지스틱 회귀분석의 적합도가 높았으며, 특히 2차 로지스틱 회귀모형을 이용한 분석이 가지고 있는 제한점인 비대칭 자료에 대해서도 확률곡선을 잘 추정할 수 있었다. 또한 이분법에 의해서 얻어지는 추정 확률함수는 1차 로지스틱 회귀모형의 합성함수 형태로 표현되므로 2차 로지스틱 회귀모형에 비해 독립변수의 효과를 설명하기도 쉽다는 것을 알 수 있었다. 결론적으로 확률곡선의 형태가 종형인 경우에 있어서 2차 로지스틱 회귀분석의 대안으로 이분법을 이용한 로지스틱 회귀분석을 사용 할 수 있으며, 이분법을 사용하는 것이 종형의 형태에 관계없이 통계적으로 유용한 확률곡선을 추정하고 결과도출이나 분석의 결과에 대한 해석에 있어서도 명확한 결과를 가져다 준다는 것을 확인하였다.

참고문헌

[1] Cox, D.R. (1970). *The Analysis of Binary Data.*, London : Methun.

[2] Alan Agresti. (1990). *An Introduction to Categorical Data.* New York : Wiley.

[3] Michael J.A. Berry, Gordon Linoff. (1997). *Data Mining Techniques.* Wiley.

[4] David W. Hosmer, Jr. Stanley Lemeshow. (1989).*Applied logistic regression.* New York : Springer-Verlag.

[5] Richard L. Burden, J. Douglas Faires. (1997). *Numerical Analysis.* Brooks/Core.

[2000년 6월 접수, 2000년 11월 채택]

Estimation of Asymmetric Bell Shaped Probability Curve using Logistic Regression*

Sung H. Park[1] Ki-ho Kim[2] So-hyung Lee[3]

ABSTRACT

Logistic regression model is one of the most popular linear models for a binary response variable and used for the estimation of probability function. In many practical situations, the probability function can be expressed by a bell shaped curve and such a function can be estimated by a second order logistic regression model. However, when the probability curve is asymmetric, the estimation results using a second order logistic regression model may not be precise because a second order logistic regression model is a symmetric function. In addition, even if a second order logistic regression model is used, the interpretation for the effect of second order term may not be easy. In this paper, in order to alleviate such problems, an estimation method for asymmetric probabiity curve based on a first order logistic regression model and iterative bi-section method is proposed and its performance is compared with that of a second order logistic regression model by a simulation study.

Keywords: Logistic regression model; Asymmetric bell shape; Estimation of probability curve.

* This work was partially supported by the Brain Korea 21 Project.

1) Professor, Department of Statistics, Seoul National University.
E-mail: parksh@plaza.snu.ac.kr
2) Head consultant, PWC Management Consulting.
E-mail: kiho.kim@kr.pwcglobal.com
3) Manager, IT Team, Foreign Exchange Card.