한국통계학회논문집 제2권 1호
1995년 4월 pp. 22-31

포아송 반응을 갖는 로그 선형 회귀 모형에 대한 최우추정량과 모의실험 연구

한정혜[1] 조중재[2]

요 약

본 논문에서는 포아송 반응을 갖는 로그 선형 회귀 모형에 붙스트랩 방법을 이용하여, 여러가지 통계적 추론을 위한 유용한 확률적 결과들을 연구 · 소개하고, 모의실험을 통한 소표본 성질들을 다양하게 제시하고자 한다. 특히 로그 선형 회귀 모형에 대한 최우추정량 $\widehat{\boldsymbol{\beta}_n}$ 및 정보행렬 $I(\boldsymbol{\beta}_0)$의 추정량들 $I_1(\widehat{\boldsymbol{\beta}_n}; \boldsymbol{X})$와 $I_2(\widehat{\boldsymbol{\beta}_n}; \boldsymbol{X})$에 대한 일치성 및 정규성등의 확률적 성질들, 그리고 붙스트랩 방법을 적용한 대표본 성질들과 관련하여 여러가지 모의실험 결과들을 분석 · 연구하였다.

1. 확률 모형과 최우추정

서로 독립이고, 동일한 분포를 갖는 확률 표본 $(Y_i, \boldsymbol{X}^t_i)$, $i=1,2,3,\cdots,n$에 대하여 공변량 $\boldsymbol{X}_i = \boldsymbol{x}_i$가 주어졌을때, 반응 변수 Y_i는 다음의 포아송 분포 함수를 따른다고 하자.

$$P(Y_i = y_i | \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\exp(-\lambda_{\boldsymbol{x}_i})\lambda^{y_i}_{\boldsymbol{x}_i}}{y_i!}$$

단, $y_i = 0,1,2,\cdots$, $\log(\lambda_{\boldsymbol{x}_i}) = \boldsymbol{x}_i^t \boldsymbol{\beta}$.

위와 같이 회귀변량 $\boldsymbol{X}_i$는 자연 링크 함수 $\theta_i = \log(\lambda_{\boldsymbol{x}_i})$를 통해 반응변수 Y_i와 연결된다. 이때 $\boldsymbol{\beta}$는 미지의 p차원 벡터로 $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^t$이고 공변량 $\boldsymbol{X}_i$는 분포함수 $G(\cdot)$를 갖고 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^t$이다. 그리고 $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20}, \cdots, \beta_{p0})^t$는 관측값들을 형성하는 미지의 실제 모수라고 하자. 그러면 한개의 관측치에 대한 로그 우도함수는 다음과 같이 주어진다.

$$l(\boldsymbol{\beta}, G; y_i, \boldsymbol{x}_i) = [\, y_i \boldsymbol{x}_i^t \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^t \boldsymbol{\beta}) - \log(y_i!)]$$

1) (360-763) 충북 청주시 흥덕구 개신동 산 48 충북대학교 전자계산학과.
2) (360-763) 충북 청주시 흥덕구 개신동 산 48 충북대학교 통계학과.

체인 법칙에 의해 로그 우도 함수 l의 β에 대한 일차, 이차 미분은 다음과 같이 비선형 함수로 주어진다.

$$\nabla l = \frac{\partial l}{\partial \beta} = x_i[\, y_i - \exp(x_i^t \beta)]$$

$$\nabla^2 l = \frac{\partial^2 l}{\partial \beta \partial \beta^t} = -[\, x_i \; x_i^t \exp(x_i^t \beta)]$$

그러므로, 주어진 표본에 대해 최우추정치 $\hat{\beta}_n$는 다음 비선형 방정식에서 구할 수 있는데 이는 수치 해석적 반복을 통해 근사적으로 유효하게 구해진다(McCullagh & Nelder, 1989).

$$\sum_{i=1}^{n} x_i(y_i - \exp(x_i^t \beta)) = 0$$

그리고 β에 대한 정보 행렬을 $I(\beta)$라 하면 이는 다음의 두 관계식으로 표현된다.

$$I(\beta) = E[\, \nabla l(\nabla l)^t] = E[\, X_i \; X_i^t \, (Y_i - \exp(X_i^t \beta))^2] \tag{1}$$

$$I(\beta) = -E(\nabla^2 l) = E[\, X_i \; X_i^t \exp(X_i^t \beta)] \tag{2}$$

따라서 두가지 표현을 기초로 다음의 두 추정행렬을 고려할 수 있을 것이다(Griffiths et al. 1987).

$$I_1(\hat{\beta}_n; X) = \frac{1}{n} \sum_{i=1}^{n} X_i \; X_i^t \exp(X_i^t \hat{\beta}_n), \qquad \text{(Newton-Raphson 알고리즘)}$$

$$I_2(\hat{\beta}_n; X) = \frac{1}{n} \sum_{i=1}^{n} X_i \; X_i^t [\, Y_i - \exp(X_i^t \hat{\beta}_n)]^2 \qquad \text{(BHHH 알고리즘)}$$

이때 $\hat{\beta}_n$은 역시 $\sum_{i=1}^{n} X_i(Y_i - \exp(X_i^t \beta)) = 0$의 해이며, 이는 반복적인 방법으로 구해진다. 그러면 적당한 조건하에서 다음 식들이 성립된다(Lee K.W. et al. 1992).

$$I_i(\hat{\beta}_n; X) \xrightarrow{w.p.1} I(\beta_0) \tag{3}$$

$$\sqrt{n} I_i^{\frac{1}{2}}(\hat{\beta}_n; X)(\hat{\beta}_n - \beta_0) \xrightarrow{d} N_p(0, I_{p\times p}) \tag{4}$$

단, $I_i^{\frac{1}{2}}(\widehat{\beta_n}; X)[I_i^{\frac{1}{2}}(\widehat{\beta_n}; X)]^t = I_i(\widehat{\beta_n}; X), \quad (i=1,2)$

2. 붙스트랩 방법과 최우추정

최근 일반화 선형 모형들에 대한 붙스트랩 연구 결과로 Lee K.W. (1990)와 Cho & Han (1994) 등이 있다. 이때, 미지의 모수벡터 β_0의 최우추정량 $\widehat{\beta_n}$과 표본 X_i에 기초한 경험적 누적 확률 분포 함수에 의해 여러가지 붙스트랩 추정량들이 다음의 절차에 의해 추정되어 진다.

1단계 : 원래의 표본 X_i와 Y_i로 부터 수치적 방법을 통해 β_0의 최우추정치 $\widehat{\beta_n}$을 계산하고, X_i에 기초한 경험적 누적 확률 분포함수를 구한다 $(i=1,2,\cdots,n)$.

2단계 : 붙스트랩 표본 X_i^*와 Y_i^*를 다음과 같은 방법에 의해 추출하자. 원래의 표본 X_i에서 복원 추출한 n개의 표본들을 X_i^*라 하자. 그리고 $X_i^* = x_i^*$가 주어졌을때, 아래의 확률분포 함수로부터 표본 Y_i^*를 추출하자 $(i=1,2,\cdots,n)$. 이때 링크 함수는 $\hat{\theta}_i^* = x_i^{*t}\widehat{\beta_n}$이다.

$$P(Y_i^* = y_i^* | X_i^* = x_i^*) = \exp[(y_i^* \hat{\theta}_{i^*} - \exp(\hat{\theta}_{i^*}) - \log(y_i^*!))]$$

3단계 : 붙스트랩 표본으로 부터 로그 우도 함수 즉, l^*를 최대로 하는 $\widehat{\beta_n^*}$을 계산한다. 다시말하면 $\widehat{\beta_n^*}$는 다음의 비선형 방정식의 해이다.

$$\sum_{i=1}^{n} X_{i^*}[Y_i^* - \exp(X_i^{*t}\beta)] = 0$$

붙스트랩 표본들을 $X^* = \left[\begin{pmatrix} Y_1^* \\ X_1^* \end{pmatrix}, \begin{pmatrix} Y_2^* \\ X_2^* \end{pmatrix}, \cdots, \begin{pmatrix} Y_n^* \\ X_n^* \end{pmatrix}\right]$ 이라하면, 적당한 조건하에서 다음과 같은 붙스트랩 버젼들이 쉽게 계산된다(Han J.H., 1994. 참조).

$$\sqrt{n}(\widehat{\beta_n^*} - \widehat{\beta_n}) \xrightarrow{d} N_p(0, I^{-1}(\beta_0)) \tag{5}$$

$$I_i(\hat{\beta}_n^*;\ X^*) \xrightarrow{p} I(\beta_0) \tag{6}$$

$$\sqrt{n} I_i^{\frac{1}{2}}(\hat{\beta}_n^*;\ X^*)(\hat{\beta}_n^* - \hat{\beta}_n) \xrightarrow{d} N_p(0, I_{p\times p}) \tag{7}$$

이때 *가 붙은 변수들은 붙스트랩 방법이 적용된 항들이다. 또한 $I_{p\times p}$는 $(p\times p)$크기인단위행렬이고, $I_i^{\frac{1}{2}}(\hat{\beta}_n^*;\ X^*)[\ I_i^{\frac{1}{2}}\ \hat{\beta}_n^*;\ X^*)]^t = I_i(\hat{\beta}_n^*;\ X^*)$, $(i=1,2)$이다.

3. 모의실험 연구

3.1 모의실험 절차

앞 절의 이론 전개와 이기원 외 2인(1993), 그리고 Griffiths et al.(1987)의 연구 결과를 고려하여 모의실험 절차를 소개하면 다음과 같다. 편의상 공변량 벡터 $X_i=(X_{i1}, X_{i2})^t$에 대하여 설명 변수 X_1, X_2가 상관계수가 ρ인 일양 분포 $U(-1,1)$을 따른다고 하자. 그리고 추정하고자 하는 미지의 모수의 참값을 $\beta_0 = (\beta_{10}, \beta_{20})^t = (0.5,1)^t$이라 가정하자. 모의실험에서 n의 값은 편의상 $n=62,122,302$을 사용했다.

우선 여러가지 상관계수 ρ값 $(0, \pm 0.3, \pm 0.8)$에 따라 $X_i=(X_{i1}, X_{i2})^t$를 n개씩 생성, 모수 $\lambda_{x_i} = \exp(\beta_{10}X_{i1} + \beta_{20}X_{i2})$을 갖는 포아송 반응값 Y_i를 n개를 생성한다. 그리고나서 β_0의 최우추정치 $\hat{\beta}_n=(\beta_{1n}, \beta_{2n})^t$을 구하기 위해 다음의 수치해석적 반복 공식

$$\beta_n^{(m)} = \beta_n^{(m-1)} + \frac{1}{n}[\ I_i(\ \beta_n^{(m-1)};\ X)]^{-1} \frac{\partial l}{\partial \beta} \Big|_{\beta=\beta^{(m-1)}}$$

을 사용하여 실험자가 임의로 정한 수렴 반경(여기서는 10^{-4})을 만족할때까지 최대 100번의 반복을 허용했다 $(i=1,2)$. 이때 $I_1(\ \beta_n^{(m-1)};\ X)$와 $I_2(\ \beta_n^{(m-1)};\ X)$은 각각 위의 수치해석적 공식에 의해 반복 계산되어진 행렬이다. 이러한 과정을 1000번 시행한다. 여기서 나온 1000개의 β_0의 추정치 $\hat{\beta}_n=(\beta_{1n}, \beta_{2n})^t$에 대한 결과들을 살펴 보았다. 여기서는 $\rho=0$인 경우만 제시하겠다. 한편, 붙스트랩 방법의 적용에 따른 모의실험은 기존의 표본들 $(Y_i, X_{i1}, X_{i2})^t$에 붙

스트랩 알고리즘을 적용하면 된다. 우선 제 2 장에서 계산된 추정치 $\widehat{\beta_n}$에 의해 붙스트랩 표본들 $(Y_i^*, X_{i1}^*, X_{i2}^*)^t$, $i=1,2,\cdots,n$을 추출하여 모수 $(\beta_{10},\beta_{20})^t$에 대한 붙스트랩 추정치 $\widehat{\beta_n^*}=(\beta_{1n}^*,\beta_{2n}^*)^t$을 앞에서와 마찬가지로 다음의 수치해석적 반복공식

$$\beta^{*(m)}=\beta^{*(m-1)}+\frac{1}{n}[\ I_i(\ \beta^{*(m-1)}:\ X^*)]^{-1}\frac{\partial l^*}{\partial\ \beta^*}\ |\ _{\beta^*=\beta^{*(m-1)}}$$

에 의해 원하는 수렴 반경(10^{-4})까지 반복,추정한다 ($i=1,2$). 이때 $I_1(\ \beta^{*(m-1)}:\ X^*)$, $I_2(\ \beta^{*(m-1)}:\ X^*)$은 붙스트랩 버젼으로 반복계산 되어진 행렬이다. 역시 이 과정을 각 n과 ρ에 대해 1000번씩 시행했다. 여기서 나온 1000개의 β_0의 붙스트랩 추정치 $\widehat{\beta_n^*}=(\beta_{1n}^*,\beta_{2n}^*)^t$에 대해서도 위와 같은 분석을 해 보았다.

3.2 모의실험 결과 분석

우선 표 1과 표 2는 모수 벡터 $\beta_0=(\beta_{10},\beta_{20})^t=(0.5,1)^t$에 대한 추정치와 이 추정치에 대한 식 (4), (5), (7)을 검정하기 위한 통계량 값들이다. 즉 $n=62,122,302$에 대하여 각각 원래 표본과 붙스트랩 표본의 추정치와 MSE값을 제시했다. n이 커짐에 따라 추정치는 모수에 더욱 근사하며 MSE값도 감소하며, 붙스트랩 추정치도 역시 같은 경향을 보인다. 또 Pr > |T|는 $H_0:\beta_i=\beta_{i0}$, $i=1,2$에 대한 t-검정 유의 확률값으로 대체로 만족스럽다. 그렇지 못한 경우는 원래 표본이 제대로 추출되지 않음으로 인해 발생된 것이다. 그리고 Pr < W는 식 (4), (5), (7)의 이변량 정규성을 보장받기 위한 필요 조건인 주변 확률의 정규성 검정의 샤피로-윌크 검정 확률값으로 모두 0.05보다 커서 역시 만족스럽다. 따라서 각 추정치들의 이변량 정규성을 시사하고 있다. 다음으로 표 3은 각 모수에 대한 일차원 정규분포 여부를 알아본 이후의 이변량 정규성 여부를 검정하는 χ^2검정값을 제시하고 있다(Johnson & Wichern, 1982. 참조). 타원 $\chi^2(2,0.05)$안에 관측되는 추정치들의 비율이 대체로 0.5를 넘어 이변량 정규성에 위배되지 않는다. 그리고 각 추정에 사용된 알고리즘이 모수 추정에 평균 몇번의 계산을 하는가를 비교한 결과, Newton-Raphson 알고리즘 정보 행렬을 사용한 경우는 평균 2~3회, BHHH 알고리즘 경우는 8~17회로 전자가 훨씬 더 효율적임을 알수 있다. 다음으로 추정치들의 정규성 적합 검정을 위해 정규 확률 분포를 등확률 25 %로 나눈 40개의 구간에 1000개의 모수 추정치에 대해 $z_i=\frac{\widehat{\beta_\epsilon}-\beta_{i0}}{SE(\widehat{\beta_\epsilon})}$ 값이 관측되는 갯수에 대하여 $Q_{39}=\sum_{i=1}^{40}\frac{(E_i-Q_i)^2}{E_i}$을 계산한 결과를 제시했다.

이 Q_{39}값이 임계치 $\chi^2(39,0.05)=55$보다 모두 작아 유의수준 $\alpha=0.05$하에서 정규 분포를 따른다고 할수 있다. 따라서 식(4), (5), (7)를 보이기 위한 표 1, 2, 3에서의 통계량 값들이 대체로 만족스러움을 보았다. 표4는 모수 추정시 사용된 알고리즘의 정보행렬 추정값과 MSE로 Mathemetica 2.0으로 계산한 $\rho=0$일때의 식 (1) 또는 (2)의 참행렬 $I(\beta)=\begin{pmatrix}0.4580 & 0.0629\\ 0.0629 & 0.4216\end{pmatrix}$ 에 대하여 주어진 표본의 크기 n에 따라 항상 Newton-Raphson 알고리즘이 보다 효율적이고, $n\to\infty$ 일때 식 (3), (6)의 일치성이 성립함을 알수있다. 그리고 표본의 크기 n, 상관계수 ρ 그리고 추정 행렬에 따른 알고리즘에 대한 각각의 추정치들의 산포도들을 분석한 결과, 그림에서 보듯 1000개의 모수 각 추정치는 중심 모수값을 중심으로 이변량 정규분포의 형태를 관찰할 수 있었다. 여기에서 제시하지는 않았지만 n이 커짐에 따라 중심 모수값에는 더욱 더 밀집되어 분포하고, 그 분산도 작아짐을, 그리고 상관계수 ρ에 따른 추정치들의 산포도 예견된 바와 같음을 알 수 있었다. 여기에서 주목할 것은 표에서 언급 되었듯이 두가지 알고리즘에 대한 추정치들의 산포는 유사하게 나타났지만, 추정치들을 얻는데 수렴에 필요한 반복 횟수나 계산 속도 측면에서 Newton-Raphson 알고리즘(I_1행렬 또는 I_1^*행렬)을 사용하는 것이 보다 효율적이라는 결론을 얻었다. 또한 붙스트랩 추정치들의 산포와 관련하여 표에서도 언급된 정규성이나 일치성 등 만족스러운 성질들을 확인할 수 있었다. 이때, 그림에서 가장 작은 타원은 이변량 정규성 검정을 위한 50%의 자유도 2인 극한 χ^2분포에 따른 신뢰 타원이며, 중간 타원과 바깥 타원은 각각 95%, 99%의 신뢰 타원이다.

<표 1> $\rho=0$일때 $\widehat{\beta_1}$

			n=62		n=122		n=302	
			original	bootstrap	original	bootstrap	original	bootstrap
$\beta_1=0.5$	$\widehat{\beta_1}$	I1	0.4995	0.4931	0.5055	0.4818	0.5007	0.5046
		I2	0.4998	0.4961	0.5047	0.4819	0.5007	0.5046
	MSE	I1	0.0413	0.0404	0.0198	0.0230	0.0084	0.0082
		I2	0.0416	0.0415	0.0199	0.0232	0.0084	0.0082
	Pr > |T|	I1	0.9375	0.2807	0.2141	0.0001	0.8164	0.1068
		I2	0.9778	0.5466	0.2896	0.0002	0.8164	0.1099
	Pr < W	I1	0.0799	0.0799	0.0055	0.0055	0.3837	0.3837
		I2	0.1270	0.1270	0.0054	0.0054	0.3837	0.3837

<표 2> $\rho=0$일때 $\widehat{\beta_2}$

			n=62		n=122		n=302	
			original	bootstrap	original	bootstrap	original	bootstrap
$\beta_2=1$	$\widehat{\beta_2}$	I1	0.9914	0.9851	0.9943	0.9845	0.9979	0.9887
		I2	0.9880	0.9830	0.9938	0.9848	0.9979	0.9887
	MSE	I1	0.0385	0.0385	0.0177	0.0214	0.0076	0.0079
		I2	0.0403	0.0370	0.0176	0.0214	0.0076	0.0079
	Pr > |T|	I1	0.1633	0.0218	0.1730	0.0008	0.4394	0.0001
		I2	0.0584	0.0050	0.1402	0.0010	0.4394	0.0001
	Pr < W	I1	0.0510	0.0510	0.0126	0.0126	0.8238	0.8238
		I2	0.0163	0.0163	0.0138	0.0138	0.8238	0.8238

<표 3> $\rho=0$일때

		n=62		n=122		n=302	
		original	bootstrap	original	bootstrap	original	bootstrap
$\chi^2(2;0.5)$내 관측비율	I1	0.4940	0.482	0.523	0.476	0.491	0.514
	I2	0.5370	0.520	0.545	0.520	0.499	0.523
평균 수렴 반복 횟수 (최소-최대)	I1	2.903 (2-4)	2.839 (1-4)	2.832 (2-4)	2.832 (2-4)	2.754 (2-3)	2.752 (2-4)
	I2	17.4339 (4-99)	17.3079 (3-99)	13.0817 (3-91)	13.8829 (3-98)	8.5706 (3-69)	8.3248 (3-40)
10^{-4} 내에 수렴 갯수	I1	1000	1000	1000	1000	1000	1000
	I2	802	825	955	948	999	999
Q_{39}	I1	36.24	37.68	43.04	39.6	38.4	45.84
	I2	38.96	43.68	49.28	54.96	43.28	58.96

<표 4> $\rho=0$

추정 정보 행렬 원소	사용 알고리즘	n=62		n=122		n=302	
		original	bootstrap	original	bootstrap	original	bootstrap
(1,1)	I1	0.4275	0.3974	0.4253	0.3758	0.4228	0.4226
	(MSE)	(0.0074)	(0.0086)	(0.0042)	(0.0092)	(0.0025)	(0.0026)
	I2	0.3971	0.3661	0.4119	0.3541	0.4166	0.4168
	(MSE)	(0.0244)	(0.0257)	(0.0145)	(0.0195)	(0.0066)	()
(1,2) 또는 (2,1)	I1	0.0641	0.0197	0.0647	-0.0322	0.5097	0.0639
	(MSE)	(0.0068)	(0.0074)	(0.0035)	(0.0120)	(0.0015)	(0.0014)
	I2	0.0565	0.0174	0.0607	-0.0390	0.0622	0.0645
	(MSE)	(0.1789)	(0.0150)	(0.0097)	(0.0181)	(0.0042)	(0.0043)
(2,2)	I1	0.4623	0.5038	0.4630	0.4373	0.4591	0.4645
	(MSE)	(0.0109)	(0.0154)	(0.0065)	(0.0042)	(0.0033)	(0.0036)
	I2	0.4315	0.4658	0.4467	0.4182	0.4522	0.4598
	(MSE)	(0.0301)	(0.0287)	(0.0182)	(0.0121)	(0.0082)	(0.0075)

참고 문헌

[1] 이기원, 손건태, 정윤식(1993). 로지스틱 회귀모형에서 최우추정량의 정확도 산정, 「응용통계연구」, 제6권 2호, 393-399.

[2] Cho J. J., Han J. H.(1994). A Study on Statistical Estimation for Generalized Logit Model of Nominal type with Bootstrap Method, *A manuscript submitted for publication.*

[3] Griffiths, W.E., Hill R.C, and Pope P.J.(1987). Small Sample Properties of Probit Model Estimators, *Journal of the American Statistical Association*, 82, 929-937.

[4] Han J.H.(1994). Bootstrapping Log-linear Model with Poisson Response and Simulation, *Unpublished manuscript.*

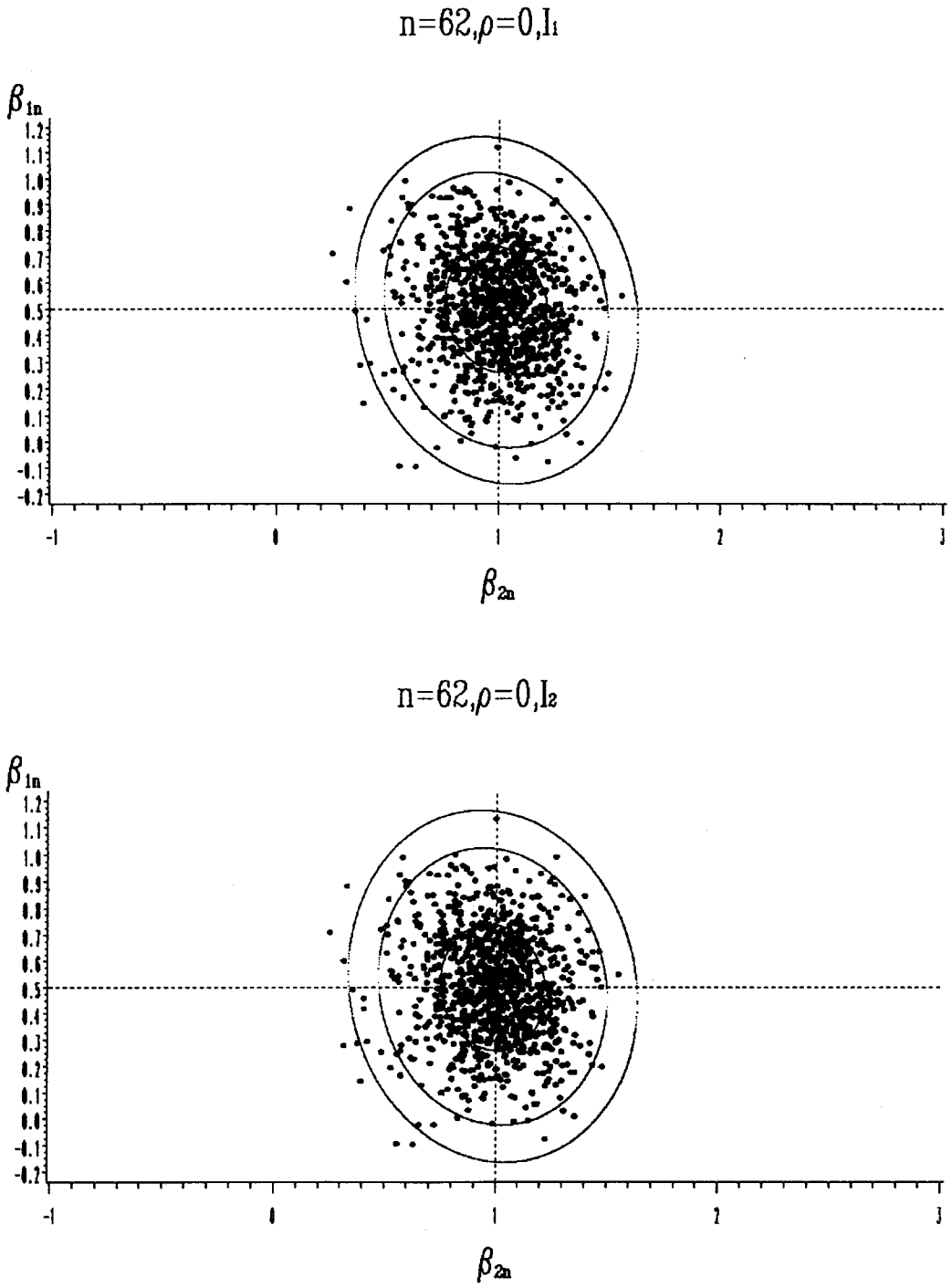
[5] Lee K.W.(1988). Bootstrap Methods in Generalized Linear Models, Ph. D. Thesis, UC. Berkeley, California 94720.

[6] Lee K.W.(1990). Bootstrapping Logistic Regression Models with Random Regressors, *Communications in Statistics(Part A : Theory and Methods)*, 19(7). 2527-2539.

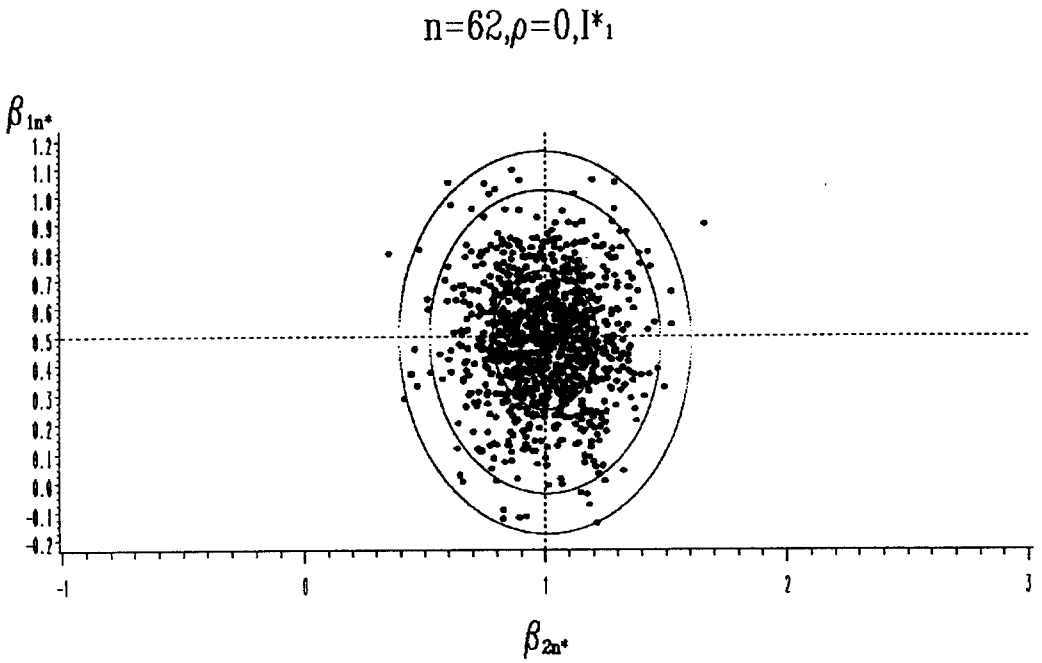
[7] Lee K.W., Kim C.R., Sohn K.T. and Jeong K.M.(1992). Bootstrapping Generalized Linear Models with Random Regressors, *Journal of the Korean Statistical Society*, 21(1). 70-79.

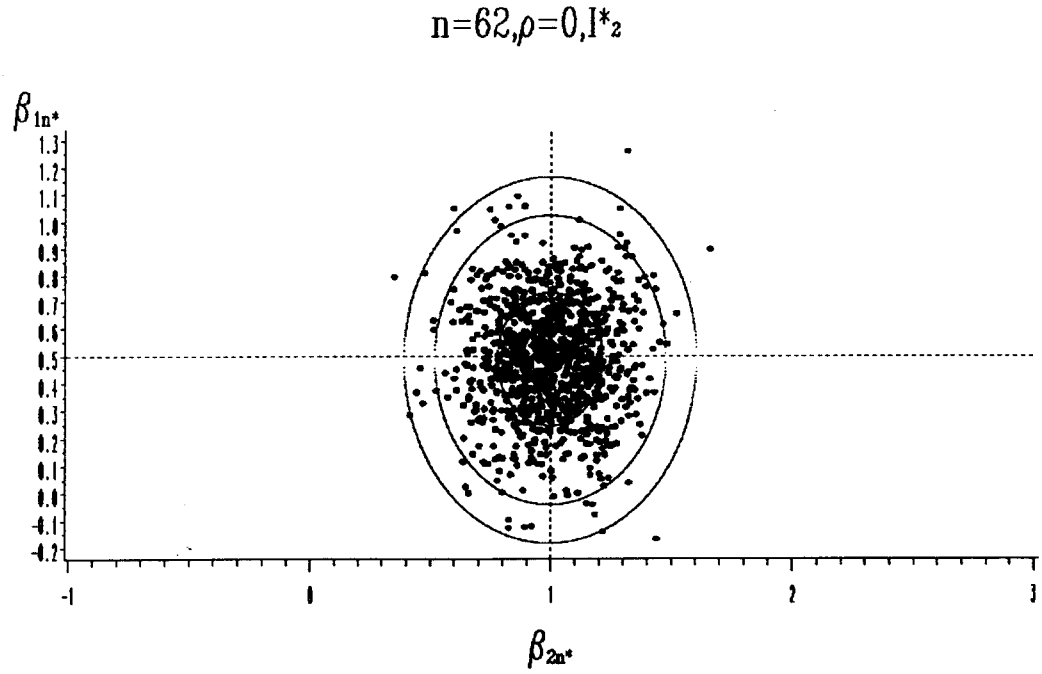
[8] McCullagh, P. and Nelder, J.A.(1989). *Generalized Linear Models*, London; Chapman and Hall.

[9] Johnson, R.A. and Wichern, D. W.(1982). *Applied Multivariate Statistical Analysis,* Prentice Hall



<그림 1> 원래표본 $n=62$, $\rho=0$





<그림 2> 붓스트랩 표본 $n=62$, $\rho=0$