응용통계연구 제 18권 2호, 2005년, pp. 435-443

수정 결정계수를 사용한 로지스틱 회귀모형에서의 변수선택법

홍종선[1] 함주형[2] 김호일[2]

요 약

로지스틱 회귀모형에서 결정계수는 선형 회귀모형보다 다양하게 정의되며 그 값들도 매우 작아 로지스틱 회귀모형 평가기준으로 사용되는 통계량이라고 할 수 없다. Liao와 McGee(2003)는 부적절한 설명변수의 추가 또는 표본크기의 변화에 민감하지 않은 두종류의 수정 결정계수를 제안하였다. 본 연구에서는 실제자료에 적용한 로지스틱 회귀모형에서 수정 결정계수를 포함한 네 종류의 결정계수들을 변수선택의 기준으로 사용하여 기존의 변수선택 방법인 전진선택, 후진제거, 단계적 선택방법, AIC 통계량 등을 사용한 방법들과 비교하여 그 적절함과 효율성을 토론한다.

주요용어: 결정계수, 로지스틱 회귀, 변수선택

1. 서론

회귀분석에서 변수선택을 위한 많은 방법들이 제안되어 왔다. 고전적인 변수선택 방법들에는 전진선택법(forward selection), 후진제거법(backward elimination), 단계적 선택법(stepwise selection) 등이 있다. 그 변수선택의 판단기준으로 다양한 통계량들이 존재하지만 결정계수(R^2)를 많이 이용한다. 로지스틱 회귀에서는 선형회귀와 달리 결정계수의 정의가 유일하지 않고 다양하기 때문에 변수선택의 판단기준으로 결정계수를 주로 사용하지 않는다.

로지스틱 회귀에서 여러 종류의 결정계수들이 Mittlbock과 Schemper(1996), Menard (2000) 등에 의해 재언급되어 왔다. 선형회귀에서의 결정계수와 같이, 로지스틱 회귀의 결정계수들 역시 연관성정도를 과대추정할 수 있다. 이에 Liao와 McGee(2003)는 선형회귀에서의 조정된 결정계수의 개념적인 확장으로서, 로지스틱 회귀에 있어서 두 종류의 수정 결정계수 통계량을 제안하였다. 제안된 수정 결정계수들은 부적절한 설명변수들의 추가 혹은 표본크기가 변화에 민감하지 않음을 보였다. 따라서 본 논문에서는 Liao와 McGee(2003)가 제안한 수정 결정계수 통계량을 변수선택의 기준으로 사용하여 기존의 여러 변수선택 방법들과 비교하고 그 효율성에 대해 살펴보는 데 연구목적이 있다.

논문의 구성은 다음과 같다. 2절에서 로지스틱 회귀에서의 다양한 결정계수들을 소개하고 선호되어 사용하는 기존의 결정계수 두 종류와 Liao와 McGee(2003)가 제안한 두 종

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
E-mail: cshong@skku.ac.kr
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 대학원생

류의 수정 결정계수들의 유의함과 그것들을 선호하는 이유를 설명한다. 3절에서는 우선 로지스틱 회귀에 있어 기존의 변수선택 방법들과 기준들을 제시한다. 그리고 제안된 결정계수들을 실제 자료에 적용하여 변수선택을 실시해 보고 기존의 변수선택 방법들과 비교분석해보았고 4절에서 이를 토의하고 결론을 유도한다.

2. 결정계수

일반적인 다변량 로지스틱 회귀모형을 고려하자.

$$z_i \sim \text{Bernoulli}(\pi_i)\ \text{logit}(\pi_i) = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip},\ i = 1, \ldots, n \tag{2.1}$$

로지스틱 회귀에서 반응변수벡터를 $z = (z_1, \ldots, z_n)$, 그리고 확률벡터를 $\pi = (\pi_1, \ldots, \pi_n)$라고 하자. 그리고 $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)$은 최대우도추정에 기초를 둔 로지스틱 회귀모형 (2.1)식 하에서 적합된 확률벡터이다.

선형 회귀모형에서 일반적으로 결정계수에 대해 인정받는 정의는 오직 하나이지만, 로지스틱 회귀모형에서는 다음과 같이 세 종류로 분류할 수 있는 다양한 결정계수들이 논의되었다(Mittlbock과 Schemper(1996), Menard(2000), Zheng과 Agresti(2000)). 먼저 z와 $\hat{\pi}$의 제곱상관계수를 이용한 결정계수들이 있다.

(1) Pearson 상관계수의 제곱 (r^2)

$$r = (\sum_{i=1}^{n} z_i \hat{\pi}_i - n\bar{\pi}^2) / \sqrt{n\bar{\pi}(1-\bar{\pi}) \sum_{i=1}^{n} (\hat{\pi}_i - \bar{\pi})^2},$$

여기서 $\bar{\pi} = \sum_{i=1}^{n} \pi_i / n$.

(2) Spearman 상관계수의 제곱 (r_s^2)

$$r_s = \sum_{i=1}^{n} (R(z_i) - \bar{R})(R(\hat{\pi}_i) - \bar{R}) / \sqrt{\sum_{i=1}^{n} (R(z_i) - \bar{R})^2 \sum_{i=1}^{n} (R(\hat{\pi}_i) - \bar{R})^2},$$

여기서 $R(z)$는 z의 순위(rank)이고 $\bar{R} = (n+1)/2$이다.

(3) Kendall의 τ_a의 제곱 (τ_a^2)

$$\tau_a = \sum_{i<j} sign(z_j - z_i) sign(\hat{\pi}_j - \hat{\pi}_i) / [n(n-1)/2],$$

여기서 $sign(z) = \begin{cases} 1 & \text{만약 } z > 0 \\ 0 & \text{만약 } z = 0 \\ -1 & \text{만약 } z < 0. \end{cases}$

(4) Kendall의 τ_b의 제곱 (τ_b^2)

$$\tau_b = \sum_{i<j} sign(z_j - z_i) sign(\hat{\pi}_j - \hat{\pi}_i) / \sqrt{\sum_{i<j} sign^2(z_j - z_i) \sum_{i<j} sign^2(\hat{\pi}_j - \hat{\pi}_i)}$$

(5) Somers의 $D_{\hat{\pi}z}$의 제곱 $(D^2_{\hat{\pi}z})$

$$D_{\hat{\pi}z} = \sum_{i<j} sign(z_j - z_i)sign(\hat{\pi}_j - \hat{\pi}_i) / \sum_{i<j} sign^2(z_j - z_i)$$

(6) Goodman과 Kruskal의 γ의 제곱 (γ^2)

$$\gamma = \sum_{i<j} sign(z_j - z_i)sign(\hat{\pi}_j - \hat{\pi}_i) / [\sum_{i<j} sign^2(z_j - z_i)sign^2(\hat{\pi}_j - \hat{\pi}_i)]$$

다음으로 z의 산포(dispersion)에서의 비례감소(proportional reduction)에 기반한 결정계수들이 있다.

(7) 제곱합 R^2 (R^2_o)

$$R^2_o = 1 - \frac{\sum_{i=1}^n (z_i - \hat{\pi}_i)^2}{\sum_{i=1}^n (z_i - \bar{z})^2}$$

(8) Gini의 집중측도 (R^2_G)

$$R^2_G = (\sum_{i=1}^n \hat{\pi}_i^2 - n\bar{\pi}^2) / [n\bar{\pi}(1 - \bar{\pi})]$$

(9) 분류오차 R^2 (R^2_{CER})

$$R^2_{CER} = [\sum_{i=1}^n D(z_i) - \sum_{i=1}^n D(z_i|x_i)] / \sum_{i=1}^2 D(z_i),$$

여기서 $D(z_i) = \begin{cases} 1 & \text{만약 } |z_i - \bar{\pi}| > 0.5 \\ 0.5 & \text{만약 } |z_i - \bar{\pi}| = 0.5 \\ 0 & \text{만약 } |z_i - \bar{\pi}| < 0.5, \end{cases}$ $\mathrm{D}(z_i|x_i) = \begin{cases} 1 & \text{만약 } |z_i - \hat{\pi}| > 0.5 \\ 0.5 & \text{만약 } |z_i - \hat{\pi}| = 0.5 \\ 0 & \text{만약 } |z_i - \hat{\pi}| < 0.5. \end{cases}$

(10) 엔트로피를 기반으로 한 R^2 (R^2_l)

$$R^2_l = 1 - \frac{l(z, \hat{\pi})}{l(z, \hat{\pi}^0)},$$

여기서 $l(z, \pi) = \sum_{i=1}^n \{z_i \log(\pi_i) + (1 - z_i)\log(1 - \pi_i)\}$로 정의되는 로그우도(log likelihood) 함수이고 $-z_i \log(\hat{\pi}_i) - (1 - z_i)\log(1 - \hat{\pi}_i)$는 엔트로피 손실(entropy loss)이다 (Efron, 1978). 그리고 $\hat{\pi}^0 = (\bar{z}, \ldots, \bar{z})$를 절편하나만을 가진 모형하에서의 적합된 확률벡터라고 하자, 여기서 $\bar{z} = \sum_{i=1}^n z_i/n$.

마지막으로 분류된 결정계수들은 우도로 나타난다.

(11) 우도비 R^2 (R^2_{LR})

$$R^2_{LR} = 1 - [L(0)/L(\hat{b})]^{2/n},$$

여기서 $L(0)$과 $L(\hat{b})$은 각각 절편하나만을 가진 모형과 적합된 모형의 우도함수이다.

(12) 수정 우도비 R^2 (R^2_{CU})

$$R^2_{CU} = R^2_{LR}/U,$$

여기서 $U = 1 - [L(0)]^{2/n}$이다.

Kvalseth(1985)는 결정계수에 대한 여덟 가지의 기준을 제시하였는데 이를 기준으로 Mittlbock과 Schemper(1996), Menard(2000)는 앞에서 제시한 12가지 결정계수들 중에서 (7)번째와 (10)번째에서 언급한 R^2_o와 R^2_l를 선호하였다. 로지스틱 회귀에서 일반적으로 사용되는 결정계수인 R^2_o와 R^2_l도 연관성 정도를 과대 추정할 수 있기 때문에 Liao와 McGee(2003)는 결정계수에 대한 다음과 같은 두 종류의 수정안을 제안하였다.

평균이 π_i인 베르누이 분포를 따르는 z_i^{new}를 로지스틱 회귀에서 z_i의 독립적인 반복이라고 하자. R^2_l와 R^2_o에서의 π_i에 의한 z_i^{new}의 내재예측오차(Inherent Prediction Error : IPE)는 다음과 같다.

$$IPE_l(\pi) \equiv n^{-1}E\{-l(z^{new},\pi)\} = -n^{-1}\sum_{i=1}^{n}\{\pi_i \log\pi_i + (1-\pi_i)\log(1-\pi_i)\}$$

$$IPE_o(\pi) \equiv n^{-1}E\sum_{i=1}^{n}\{(z_i^{new}-\pi_i)^2\} = n^{-1}\sum_{i=1}^{n}\pi_i(1-\pi_i)$$

참의 π_i가 z_i^{new}를 예측하는 데 이용되기 때문에 이 오차를 내재예측오차라 부른다. 다음으로 로지스틱 모형에서의 추정량들을 고려하면 $IPE_l(\pi)$와 $IPE_o(\pi)$의 단순한 추정량은 각각 $-n^{-1}l(z,\hat{\pi})$과 $n^{-1}\sum_{i=1}^{n}(z_i-\hat{\pi}_i)^2$이다. 두 추정량은 $\hat{\pi}$과 z의 거리가 π와 z의 거리보다 줄어드는 경향이 있기 때문에 0에 가까운 편의를 가지고 있다. 그 편의들은 각각 다음과 같다.

$$B_l(\pi) = n^{-1}E\{-l(z^{new},\hat{\pi}^{new})\} - IPE_l(\pi)$$

$$B_o(\pi) = n^{-1}E\sum_{i=1}^{n}\{(z_i^{new},\hat{\pi}_i^{new})^2\} - IPE_o(\pi),$$

여기서 $\hat{\pi}_i^{new}$은 z_i^{new}에 대한 다변량 로지스틱 회귀모형 (2.1)식에서의 적합된 확률이다. 이후에 선형회귀에서의 R^2_{adj}를 고려하여 로지스틱 회귀에 대한 수정 결정계수들을 다음과 같이 정의한다.

$$R^2_{l,adj} = 1 - \frac{\widehat{IPE}_l^{\,p}}{\widehat{IPE}_l^{\,0}}$$

$$R^2_{o,adj} = 1 - \frac{\widehat{IPE}_o^{\,p}}{\widehat{IPE}_o^{\,0}},$$

여기서 $\widehat{IPE}_l^{\,p} = -n^{-1}l(z,\hat{\pi}) - B_l(\hat{\pi})$, $\widehat{IPE}_o^{\,p} = n^{-1}\sum_{i=1}^{n}(z_i-\hat{\pi}_i)^2 - B_o(\hat{\pi})$이고 절편하나만을 갖는 모형 하에서 $IPE(\pi)$의 편의 수정된 추정량들 $\widehat{IPE}_l^{\,0}$과 $\widehat{IPE}_o^{\,0}$은 적합된 모형에서의 $\hat{\pi}$과 $\hat{\pi}^{new}$를 제외한 같은 공식을 사용하여 얻을 수 있다.

Liao와 McGee(2003)가 제안한 수정 결정계수 $R^2_{l,adj}$와 $R^2_{o,adj}$들은 R^2_l와 R^2_o를 수정 보완하였기에 그 유의함을 그대로 계승하였고 Kvalseth(1985)의 여덟 가지 기준에도 만족스럽다. 또한 모의실험을 바탕으로 R^2_l와 R^2_o보다 부적절한 설명변수의 추가 혹은 표본크기 변화에 민감하지 않다는 결론을 내렸다(Liao와 McGee, 2003). Liao와 McGee(2003)가 제안한 수정 결정계수들을 R로 작성하여 모의실험을 했고, 그 함수들은 http://www.geocities.com/jg_liao /software에서 이용할 수 있다.

3. 최대결정계수 선택법

3.1. 변수선택법과 기준

로지스틱 회귀모형의 적합도를 평가하기 위하여 많이 이용되는 척도는 score χ^2 검정통계량과 Akaike(1973)가 제안한 AIC(Akaike's Information Criterion) 등이 있고 전진선택, 후진제거, 단계적 선택, 그리고 결정계수 R^2_l와 R^2_o를 사용한 최대결정계수 선택법으로 변수선택을 실시한다. 이러한 기존의 변수선택 기준과 더불어 수정 결정계수 $R^2_{l,adj}$와 $R^2_{o,adj}$들을 기준으로 최대결정계수 선택법을 실시하였고 기존의 변수선택 방법들과 비교해 보고자 한다.

score χ^2 통계량은 로그우도에 대한 편미분의 비율로 구하며, Akaike(1973)가 제안한 변수선택기준인 AIC는 다음과 같은 식으로 구한다.

$$AIC = -2l(z, \hat{\pi}) + 2(p+1),$$

여기서 p는 설명변수의 수이다. 이 식에서 적용자료의 표본의 크기(n)가 모두 같은 경우이기 때문에 n을 고려하지 않았다. AIC, R^2_l, R^2_o, $R^2_{l,adj}$, $R^2_{o,adj}$들은 통계프로그램인 R을 통해 계산했고 score χ^2는 SAS를 통해 구했다. 전진선택, 후진제거, 단계적 선택법은 SAS를 이용하였다. 전진선택법으로 유의한 설명변수를 추가 선택할 때에 사용되는 유의수준(SLE)을 0.05로 설정했고 후진선택법에서 변수제거시 유의수준(SLS)을 0.10, 단계적 선택법에서는 추가 혹은 제거하기 위해 SLE=SLS=0.15를 지정했다. 이러한 유의수준들은 SAS에서 임의적으로 지정해줄 수는 있지만 일반적으로 사용하는 고정된 값(default)선택하였다. AIC, R^2_l, R^2_o, $R^2_{l,adj}$, $R^2_{o,adj}$, score χ^2 통계량 등을 기준으로 한 최량선택법(best subset selection)의 경우 설명변수의 개수에 따라 가능한 모든 모형을 고려한 후에 변수의 수가 같은 단계마다 최량값(AIC 통계량은 최소값, 결정계수와 score χ^2 통계량은 최대값)과 이 값에 대응하는 모형에 포함된 변수들을 표에서 제시해 주었다.

3.2. 적용 자료 1

수정 결정계수들을 기준으로 최대결정계수 선택법을 적용해본 첫번째 자료는 Brown (1980)의 전립선암 자료이다. 이 자료는 전립선암 환자 53명에 대해 조사된 설명변수들과 림프절의 암에 대한 양성여부(Y)가 어떠한 관계에 있는지에 초점을 맞춘다. 설명변수는 X선결과(X_1), 질병단계(X_2), 종양의 등급(X_3), 환자의 연령(X_4), 혈청인산염(X_5)이다. 표 3.1은 이 자료에 대한 AIC, R^2_l, R^2_o, $R^2_{l,adj}$, $R^2_{o,adj}$, score χ^2 통계량 등을 기준으로 한 최량변

수선택법의 결과이며, 변수선택법을 사용하여 구한 최종모형에 대한 결과는 표 3.2에 나열하였다.

표 3.1: 전립선암 자료에 적용한 최량변수선택 결과

AIC		R^2					score χ^2	
선택 변수	값	선택 변수	R_l^2	R_o^2	$R_{l,adj}^2$	$R_{o,adj}^2$	선택 변수	값
1	63.008	1	0.160	0.213	0.142	0.197	1	11.283
1 2	59.353	1 2	0.241	0.302	0.205	0.270	1 2	15.714
1 2 5	**58.660**	1 2 5	0.279	**0.334**	**0.224**	**0.283**	1 2 5	17.696
1 2 4 5	59.097	1 2 4 5	**0.301**	0.346	0.236	0.286	1 2 4 5	**18.837**
1 2 3 4 5	60.126	1 2 3 4 5	0.315	0.362	0.229	0.281	1 2 3 4 5	19.451

표 3.2: 전립선암 자료에 적용한 전진·후진·단계적 선택법 결과

	전진선택법(SLE=0.05)		후진제거법(SLS=0.10)		단계적 선택법(SLE=SLS=0.15)		
step	추가	score χ^2(p값)	제거	wald χ^2(p값)	추가, 제거	score χ^2(p값)	wald χ^2(p값)
1	x1	11.283(0.001)	x3	0.976(0.323)	x1 추가	11.283(0.001)	
2	x2	5.639(0.018)	x4	1.497(0.221)	x2 추가	5.639(0.018)	
3			x5	2.660(0.103)			
최종모형	**x1 x2**		**x1 x2**		**x1 x2**		

먼저 AIC를 변수선택 기준으로 살펴보면 값이 감소했다가 증가했고, 설명변수의 수가 3개일 때가 가장 작은 값(58.660)을 가지며 이때 선택된 설명변수가 X_1, X_2, X_5인 최종모형이었다. R_l^2, R_o^2, score χ^2 통계량들의 최대값은 계속해서 증가하는데 R_l^2, score χ^2 통계량인 경우에는 모형에 포함된 설명변수들의 수가 4개(X_1, X_2, X_4, X_5), R_o^2 통계량인 경우에는 3개(X_1, X_2, X_5)인 경우의 값들이 전 단계까지는 증가폭이 컸으나 다음 단계 값들의 증가폭은 작았다. 따라서 R_l^2, score χ^2 통계량들을 기준으로 변수선택한 결과 설명변수가 X_1, X_2, X_4, X_5인 최종모형으로 선택했으며 R_o^2의 경우에는 X_1, X_2, X_5를 선택했다. 수정결정계수 $R_{l,adj}^2$, $R_{o,adj}^2$ 통계량의 값들은 설명변수의 개수가 증가해도 지속적으로 증가하지 않으며, 증가하다가 감소하는 경향이 있다. 이는 결정계수의 과대추정을 수정해준 것이며, 결정계수 값의 변화폭이 가장 작은 경우인 설명변수의 수가 3개(X_1, X_2, X_5)인 모형을 최종모형으로 선택하며 이때의 값은 각각 0.224와 0.283이다. 전진, 후진, 단계적 선택법에 의한 최종모형들은 모두 모형에 포함된 설명변수들이 2개(X_1, X_2)인 경우였다. 설명변수가 X_1, X_2인 최종모형은 변수를 선택하거나 제거하는 기준이 되는 유의수준들(SLE, SLS)을 다르게 지정하면 달라질 수 있지만 이미 설명했듯이 고정된(default) 유의수준을 사용한 결과이다.

전진, 후진, 단계적 변수선택법에 의해 구한 최종모형의 설명변수는 두 개이며 X_1과 X_2이었으나, score χ^2과 R_l^2 통계량에 의한 최종모형의 설명변수는 네 개이며 X_1, X_2, X_4, X_5이었다. AIC 통계량과 R_o^2, $R_{l,adj}^2$, $R_{o,adj}^2$ 통계량에 의한 최종모형의 설명변수는 세 개이

표 3.3: 기업파산 자료에 적용한 최량변수선택 결과

AIC		R^2					score χ^2	
선택 변수	값	선택 변수	R_l^2	R_o^2	$R_{l,adj}^2$	$R_{o,adj}^2$	선택 변수	값
3	39.344	3	0.443	0.577	0.420	0.562	3	17.336
1 3	**34.658**	1 3	0.548	0.658	**0.502**	**0.624**	2 3	21.566
1 3 4	35.533	1 3 4	**0.566**	**0.673**	0.497	0.621	2 3 4	**22.237**
1 2 3 4	37.467	1 2 3 4	0.567	0.672	0.479	0.604	1 2 3 4	22.304

표 3.4: 기업파산 자료에 적용한 전진·후진·단계적 선택법 결과

	전진선택법(SLE=0.05)		후진제거법(SLS=0.10)		단계적 선택법(SLE=SLS=0.15)		
step	추가	score χ^2(p값)	제거	wald χ^2(p값)	추가, 제거	score χ^2(p값)	wald χ^2(p값)
1	x3	17.336(< .000)	x2	0.066(0.797)	x3 추가	17.336(< .000)	
2	x1	5.981(0.015)	x4	0.975(0.324)	x1 추가	5.981(0.015)	
최종모형	**x1 x3**		**x1 x3**		**x1 x3**		

며 X_1, X_2, X_5으로 앞에서와 상이한 최종모형을 유도하였다. 전립선암 자료를 여러 변수선택방법을 사용하여 분석해 본 결과, 수정 결정계수 $R_{l,adj}^2$와 $R_{o,adj}^2$를 사용한 최대결정계수 선택법에 의하여 선정된 최종모형과 가장 보편적이고 일반적으로 많이 사용하는 AIC 통계량 기준에 의한 변수선택법의 결과가 같았던 점은 주목할 만한 부분이다.

3.3. 적용 자료 2

두 번째 적용 자료는 성웅현(2001)의 기업파산에 관한 것이다. 파산 선고된 21개 기업과 정상적으로 운영되는 25개 기업에 대해 각각 현금흐름 대 총부채비율(X_1), 순이익 대 총자산비율(X_2), 유동자산 대 유동부채비율(X_3), 유동자산 대 순매출액비율(X_4)을 조사했다. 이 자료에 대한 변수선택의 요약된 결과는 표 3.3과 표 3.4에서 보여주었다.

변수선택 방법이나 기준은 전립선암 자료의 경우와 같으므로 결과만 살펴보기로 한다. 우선 AIC, R_l^2, $R_{l,adj}^2$, $R_{o,adj}^2$, score χ^2 통계량들은 이전과 동일한 움직임을 보였다. 그러나 R_o^2는 이전과 달리 마지막 단계에서 감소하였으나 그 값의 변동이 작기 때문에 무시한다. 선택된 최종모형은 R_l^2와 R_o^2 통계량을 기준으로 했을 때 설명변수가 3개(X_1, X_3, X_4), score χ^2 통계량 기준에서 설명변수가 3개(X_2, X_3, X_4)인 경우를 제외하고는 나머지 방법은 설명변수가 2개(X_1, X_3)인 경우로 동일했다. 그리고 수정 결정계수 $R_{l,adj}^2$와 $R_{o,adj}^2$를 이용한 방법의 결과가 다른 방법들, 특히 AIC에 의해 변수선택을 실시한 결과와 동일하였다. AIC 값은 최종적으로 선택된 설명변수가 2개인 X_1, X_3일 때에 최량값 34.658를 가졌다. 이와 동일한 변수들이 선택된 모형의 $R_{l,adj}^2$와 $R_{o,adj}^2$ 통계량값들은 각각 0.502, 0.624이었고 이후 단계의 통계량들은 감소했다. 이는 Brown(1980)의 전립선암 자료의 경우에서도 유사하게 이끌어 낼 수 있는 결과이다. 따라서 변수선택기준으로 R_l^2와 R_o^2를 사용하는 것보다 수정 결정계수 $R_{l,adj}^2$와 $R_{o,adj}^2$를 사용하여 최대결정계수 선택법을 실시하는 것이 더 양호하다는

것을 유도할 수 있다.

4. 결론

Kvalseth(1985)가 제시한 로지스틱회귀에서의 결정계수에 관한 여덟 가지 기준을 바탕으로, 여러 종류의 결정계수들 중에서 Mittlbock과 Schemper(1996) 그리고 Menard(2000)가 선호한 R_o^2와 R_l^2를 가장 많이 사용하고 있다. Liao와 McGee(2003)는 부적절한 예측변수의 추가와 표본크기의 증감에 민감하지 않은 두 종류의 수정 결정계수 $R_{o,adj}^2$와 $R_{l,adj}^2$를 제안하였는데, 이러한 네 종류의 결정계수를 사용한 최대결정계수 선택법과 기존의 여러 변수선택 방법들과 실증적인 예제를 이용하여 자료에 적합한 최종모형을 비교해 보았다.

변수선택 기준이 되는 AIC, score χ^2 통계량을 사용하는 방법과 전진선택, 후진제거, 단계적 선택방법들 그리고 결정계수 R_o^2, R_l^2, $R_{o,adj}^2$, $R_{l,adj}^2$를 이용한 변수선택법으로 구한 최종모형이 대부분의 자료에서는 동일하였지만, 일부의 자료에서는 상이한 결과를 제공하였다. 이런 경우에는 수정 결정계수 $R_{o,adj}^2$와 $R_{l,adj}^2$를 이용한 최대결정계수 선택법의 결과가 AIC 통계량을 이용하여 구한 최종모형의 결과와 일치하는 경향이 있음을 발견하였다. 또한 여러 방법들의 결과와 비교적 일치하는 결과를 제시하는 결정계수는 Liao와 McGee(2003)가 제안한 수정 결정계수를 사용하는 최대결정계수 선택법임을 살펴보았다.

참고문헌

성웅현 (2001). <응용 로지스틱 회귀분석>, 탐진, 서울.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *In Proceedings of the 2nd International Symposium on Information theory*, edited by B. N. Petrov and F. Czaki. Akademiai Kiado, Budapest.

Brown, B. W. (1980). *Prediction analyses for binary data*, Biostatistics Casebook, John Wiley and Sons, New York.

Efron, B. (1978). Regression and ANOVA With Zero-One Data: Measures of Residual Variation, *Journal of the American Statistical Association*, **73**, 113-121.

Kvalseth, T. O. (1985). Cautionary Note About R^2, *The American Statistician*, **39**, 279-285.

Liao, J. G. and McGee, D. (2003). Adjusted Coefficients of Determination for Logistic Regression, *The American Statistician*, **57**, 161-165.

Menard, S. (2000). Coefficients of Determination for Multiple Logistic Regression Analysis, *The American Statistician*, **54**, 17-24.

Mittlbock, M. and Schemper, M. (1996). Explained Variation for Logistic Regression, *Statistics in Medicine*, **15**, 1987-1997.

Zheng, B. and Agresti, A. (2000). Summarizing the Predictive Power of a Generalized Linear Model, *Statistics in Medicine*, **19**, 1771-1781.

[2004년 7월 접수, 2004년 12월 채택]

Variable Selection for Logistic Regression Model Using Adjusted Coefficients of Determination

C. S. Hong[1] J. H. Ham[2] H. I. Kim[2]

ABSTRACT

Coefficients of determination in logistic regression analysis are defined as various statistics, and their values are relatively smaller than those for linear regression model. These coefficients of determination are not generally used to evaluate and diagnose logistic regression model. Liao and McGee (2003) proposed two adjusted coefficients of determination which are robust at the addition of inappropriate predictors and the variation of sample size. In this work, these adjusted coefficients of determination are applied to variable selection method for logistic regression model and compared with results of other methods such as the forward selection, backward elimination, stepwise selection, and AIC statistic.

Keywords: Coefficient of determination, Logistic regression, Variable selection.

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745 Korea.
E-mail: cshong@skku.ac.kr
2) Graduate Student, Department of Statistics, Sungkyunkwan University, Seoul 110-745 Korea.