Journal of the Korean Data & Information Science Society
2013, **24**(1), 125–133

http://dx.doi.org/10.7465/jkdi.2013.24.1.125
한국데이터정보과학회지

호흡곤란 환자 퇴원 결정을 위한 벌점 로지스틱 회귀모형

박철용[1] · 계묘진[2]

[12]계명대학교 통계학과

접수 2012년 12월 14일, 수정 2013년 1월 5일, 게재확정 2013년 1월 10일

요 약

이 논문에서는 호흡곤란을 주호소로 내원한 668명의 환자를 대상으로 11개 혈액검사 결과를 이용하여 퇴원여부를 결정하는 벌점 이항 로지스틱 회귀 기반 통계모형을 유도하였다. 구체적으로 L^2 벌점에 근거한 능형 모형과 L^1 벌점에 근거한 라소 모형을 고려하였다. 이 모형의 예측력 비교 대상으로는 일반 로지스틱 회귀의 11개 전체 변수를 사용한 모형과 변수선택된 모형이 사용되었다. 10-묶음 교차타당성 (10-fold cross-validation) 비교 결과 능형 모형의 예측력이 우수한 것으로 나타났다.

주요용어: 교차타당성, 벌점 로지스틱 회귀, 퇴원 결정, 호흡곤란 환자.

1. 머리말

호흡곤란은 응급실에서 흔히 볼 수 있는 주호소 (chief complaint) 중 하나이다. 이러한 호흡곤란의 원인질환은 짧은 시간의 문진으로 진단을 하기 어려우며, 임상전문가들은 혈액검사나 흉부 방사선 검사 등을 이용하여 진단을 하고 있다. Park 등 (2010)에서는 호흡곤란 환자의 퇴원 결정을 위한 11개 혈액검사 결과에 기반한 간편한 통계모형을 제안하였다. 또한 Park (2011)에서는 수량화 기법을 이용하여 11개 혈액검사의 중요성을 고려한 통계모형을 제안하였다. 이 방법은 결국 Fisher의 선형판별함수 (linear discriminant function)에 근거한 통계모형이 된다.

이 연구는 혈액검사 결과에 근거하여 호흡곤란 환자의 퇴원여부 예측력을 더욱 높이려는 시도에서 시작되었다. 구체적으로 호흡곤란을 주호소로 내원한 환자를 대상으로 11개 혈액검사 결과를 이용하여 환자의 퇴원여부를 결정하는 벌점 로지스틱 회귀모형 (penalized logistic regression model) 기반 통계모형을 유도하고자 한다.

Tibshirani (1996)에 의해 Lasso (least absolute shrinkage and selection operator) 기법이 소개된 이후 벌점 회귀모형 (penalized regression model)에 대한 연구가 상당히 진전되게 되었다. 기존의 능형 회귀 (ridge regression)가 L^2 벌점 (penalty)에 근거한 기법인데 반해 Lasso는 L^1 벌점에 근거한 기법이다. Tibshirani (1996)에 의하면 Lasso 회귀는 능형 회귀와는 달리 회귀계수 추정값이 0인 모형이 보다 쉽게 선택되기 때문에 변수선택의 효과가 있다고 하였다. 따라서 Lasso 회귀는 능형회귀의 장점인 예측정확도 (prediction accuracy)과 변수선택 회귀의 장점인 해석력 (interpretation)을 어느 정도 겸비할 수 있는 장점이 있다고 할 수 있다.

Tibrashini (1996)에 의해 회귀모형에 대한 Lasso 방법이 소개된 이후 여러 분야에서 Lasso 방법이 사용되고 있다. 예를 들어 일반화선형모형 (generalized linear model)과 다변량분석 (multivariate

[1] 교신저자: (704-701) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과, 교수.
E-mail: cypark1@kmu.ac.kr
[2] (704-701) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과, 석사과정생.

analysis)에서 Lasso 방법의 연구가 활발하게 진행되고 있다. 일반화선형모형에서의 연구로는 Friedman 등 (2008)을 비롯한 여러 연구가 있으며, 다변량분석의 일종인 선형판별분석 (linear discriminant analysis)에 대한 연구의 예로는 Whitten과 Tibshirani (2011) 등의 연구가 있다.

이 연구에서 고려하고 있는 통계모형은 Lasso와 능형이 포함된 벌점 로지스틱 회귀모형이다. 구체적으로 벌점 회귀모형의 잔차제곱합 위치에 -(로지스틱 회귀모형의 로그우도)를 대체한 것이 벌점 로지스틱 회귀모형이다. 벌점 로지스틱 회귀모형과 비교대상이 될 모형으로는 일반 로지스틱 회귀의 11개 혈액검사 전체를 사용한 모형과 변수선택에 의한 모형이 될 것이다. 우선적으로 예측정확도에 근거하여 최적의 모형을 선택할 것이며, 해석력 관점에서도 간단하게 논의될 것이다.

이 연구에서는 혈액검사의 원자료를 사용하지 않는다. 왜냐하면 11개 혈액검사 모두에서 정상인 내원환자들이 중앙값 근처에 몰려 있을 것이기 때문이다. 다시 말해 혈액검사 결과값이 커짐에 따라 정상적인 환자일 가능성이 단조증가하거나 단조감소하는 형태가 나오지 않기 때문에 원자료에 의한 로지스틱 회귀모형은 예측력이 높지 않을 가능성이 아주 높기 때문이다. 따라서 이 연구에서는 Park 등 (2010)에 의해 설정된 (혈액검사 결과들의) 퇴원구간에 근거한 이분형 변수를 사용한다. 이 퇴원구간은 해당 혈액검사 결과의 퇴원환자 상대도수가 퇴원환자의 상대도수보다 높은 구간으로, 퇴원구간에 속하게 되면 퇴원환자일 가능성이 높아지게 된다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 Park 등 (2010)에 소개된 연속형 변수의 이산화 방법과 함께 벌점 로지스틱 회귀모형을 간략히 설명한다. 3절에서는 호흡곤란 환자에 대해 벌점 로지스틱 회귀모형을 적합하는 과정을 설명한다. 또한 벌점 로지스틱 회귀모형을 일반 로지스틱 회귀모형을 이탈도 (deviance) 관점에서 비교한다. 4절의 결론 및 논의에서는 이 연구의 결과들을 요약하고 논점을 정리한다.

2. 이산화 방법과 벌점 로지스틱 회귀모형

이 절에서는 Park 등 (2010)에서 제시된 연속형 변수의 이산화 방법을 먼저 간략히 소개한다. 또한 이 연구에서 사용하고자 하는 벌점 이항 로지스틱 회귀모형과 함께 여러 로지스틱 회귀모형을 비교하기 위한 기준 (criterion)을 소개한다.

Park 등 (2010)에서 제안하는 이산화 방법은 우도비 (likelihood ratio)에 근거한 방법으로, 두 모집단 A (admission; 입원), D (discharge; 퇴원)의 확률밀도함수를 각각 $f_A(x), f_D(x)$라고 했을 때 다음과 같다.

$$f_D(x)/f_A(x) > 1\text{이면 } x\text{는 } D\text{집단으로 분류한다.} \tag{2.1}$$

식 (2.1)에 근거한 퇴원구간 (discharge interval)을 $\{x : f_D(x)/f_A(x) > 1\}$로 잡을 수 있다. 이 방법은 두 모집단의 사전확률이 동일할 경우 오분류확률 (misclassification probability)을 최소화하는 규칙이 된다 (Johnson과 Wichern, 1992).

식 (2.1)은 모집단에 근거한 방법이기 때문에 표본 $x_1, x_2, \ldots, x_n$에 근거한 방법을 제시한다. 확률밀도함수의 추정량으로는 다음과 같은 커널밀도함수 (kernel density function)를 이용한다.

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left[(x - x_i)/h\right]. \tag{2.2}$$

여기서 $K(x)$는 표준정규 커널밀도함수이다. 이 때 평활모수 (smoothing parameter)로는 표준정규 커널밀도함수에 적합하다고 알려져 있는 다음의 식을 사용한다.

$$h = [4/(3n)]^{1/5} S. \tag{2.3}$$

여기서 S는 $x_1, x_2, \ldots, x_n$의 표준편차이다. Park 등 (2010)에서 이산화 방법으로 제안하는 규칙을 하나의 식으로 정리하면 다음과 같다.

$$\widehat{f}_{Dh_1}(x)/\widehat{f}_{Ah_2}(x) > 1\text{이면 } x\text{는 } D\text{집단으로 분류한다.} \tag{2.4}$$

여기서 $\widehat{f}_{Dh_1}(x)$ $\widehat{f}_{Ah_2}(x)$는 식 (2.2)에 의해 퇴원과 입원 집단에서 각각 구해진 커널밀도함수이며, h_1 h_2는 각각의 집단에서 식 (2.3)에 의해서 구해진 것이다.

자료기반 이산화 기법으로는 카이제곱 방법 (Kerber, 1992), 엔트로피 방법 (Fayyad와 Irani, 1993) 및 분포기반 기법 (Lee 등, 2003) 등이 있는데 각기 나름의 단점이 있다. 카이제곱 방법과 엔트로피 방법은 추가 매개변수를 추정해야 하는 약점이 있으며, 분포기반 기법은 커널밀도함수와 같은 정밀한 확률밀도함수 추정량을 이용하지 않고 있는 단점이 있다. 참고로 Na 등 (2005)와 Kim 등 (2005)에서는 여러 이산화 알고리즘을 비교하는 연구를 수행하였다.

다음으로 벌점 (이항) 로지스틱 회귀모형에 대해서 간략히 설명한다. 벌점 (이항) 로지스틱 회귀모형에서는 다음의 목적함수 (objective function)를 최소화시키는 벌점 회귀계수 $\hat{\underline{\beta}}^P$를 추정하고자 한다.

$$\begin{aligned}\hat{\underline{\beta}}^P &= \text{argmin}_{\underline{p}(\underline{\beta})}\left\{-\ell(\underline{p}(\underline{\beta});\underline{y}) + P_\lambda^k(\underline{\beta})\right\} \\ &= \text{argmin}_{\underline{\beta}}\left\{-\sum_{i=1}^{n}\left[y_i \log p_i(\underline{\beta}) + (1-y_i)\log[1-p_i(\underline{\beta})]\right] + P_\lambda^k(\underline{\beta})\right\}.\end{aligned} \tag{2.5}$$

여기서 $\ell(\underline{p}(\underline{\beta});\underline{y})$는 로지스틱 회귀모형의 로그우도 함수, $\underline{y} = (y_1, \ldots, y_n)^T$는 0과 1의 값을 가지는 이항 반응변수 벡터이며 $\underline{p}(\underline{\beta}) = (p_1(\underline{\beta}), \ldots, p_p(\underline{\beta}))^T$는 $\underline{y}$가 1이 될 조건부 확률이며 $P_\lambda^k(\underline{\beta}) = \lambda\sum|\beta_i|^k$는 벌점 함수 (penalty function)이다. 구체적으로 $p_i(\underline{\beta}) = \exp(\alpha + \underline{\beta}^T\underline{x}_i)/[1+\exp(\alpha+\underline{\beta}^T\underline{x}_i)]$로서 $\underline{x}_i = (x_{i1}, \ldots, x_{ip})^T$는 i-번째 개체의 설명변수 벡터이다. 또한 $P_\lambda^k(\underline{\beta})$는 k가 1이면 Lasso, 2이면 능형 벌점 함수가 되는 것을 알 수 있다.

회귀계수 β_i 값은 설명변수의 척도 (scale)에 의존하기 때문에 설명변수를 사용하기 전에 사전적으로 표준화시키는 것이 필요하다. 이렇게 표준화된 설명변수를 사용하면 별점 회귀모형, 즉 목적함수가 $\sum_{i=1}^n(y_i - \alpha - \underline{\beta}^T\underline{x}_i)^2 + P_\lambda^k(\underline{\beta})$인 경우에는 $\hat{\alpha} = 0$가 되지만, 벌점 로지스틱 회귀모형에서는 $\hat{\alpha} = 0$이 항상 성립하는 것은 아니기 때문에 절편을 모형에 포함시킨다.

마지막으로 여러 가지 로지스틱 회귀모형의 성능을 비교하기 위해서 사용할 측도인 이탈도 (deviance)를 소개한다. 로지스틱 회귀모형에 대한 이탈도는 다음과 같이 정의된다 (McCullogh와 Nelder, 1989).

$$D = 2\left(\ell(\underline{y};\underline{y}) - \ell(\underline{p}(\hat{\underline{\beta}});\underline{y})\right). \tag{2.6}$$

여기서 $\ell(\underline{y};\underline{y})$는 $\underline{p}(\underline{\beta}) = \underline{y}$인 경우의 완전 (full) 로그우도 함수 값으로서 0이 되며, $\ell(\underline{p}(\hat{\underline{\beta}});\underline{y})$는 현재 고려 중인 모형의 로그우도 함수 값이다. 따라서 로지스틱 회귀모형에서의 이탈도는 고려 중인 모형의 로그우도 함수 값에 -2배를 해주면 구할 수 있게 된다.

3. 호흡곤란 환자 자료에의 적용

3.1. 설명변수의 이산화 과정

이 연구에서 사용된 자료는 모 의료원에 2006년 7월부터 2007년 6월 사이에 호흡곤란을 주호소로 내원한 환자 1129명의 의무기록에서 추출되었다. 이렇게 추출된 자료 중 타병원으로 옮긴 환자, 도착 직후 사망, 심폐소생술 후 혹은 심폐소생술 금지로 사망한 환자, 자의 퇴원 또는 미상의 기타 환자, 의무

기록이 불완전한 경우 등을 제외한 668명의 환자의 자료가 이용되었다. 이 중 500명이 입원 환자였으며 나머지 168명이 퇴원 환자였다. 또한 원래 의무기록에서 추출된 55개의 변수 중 임상전문가에 의해 중요하다고 판단된 11개의 혈액검사 변수를 분석에 사용하였다. 구체적으로 이 연구에서 사용된 11개의 설명변수는 다음의 Table 3.1에 주어져 있다.

Table 3.1 Explanatory variables used in analysis

Variable name	Explanation	Unit
WBC	White Blood Cell [count]	$\times 10^3/uL$
PLT	Platelet count	$\times 10^3/uL$
Cl-	Chloride	$mmol/L$
AST	Aspartate Transaminase	U/L
ALT	Alanine Transaminase	U/L
$PCO2$	Pressure of Carbon dioxide	$mmHg$
$PO2$	Pressure of Oxygen	$mmHg$
$O2SAT$	Oxygen Saturation	%
LDH	Lactate Dehydrogenase	U/L
$Ca2+$	Calcium	mEg/L
$Mg2+$	Magnesium	mEg/L

상대도수에 근거한 이산화 방법인 분류규칙 (2.4)를 적용한 결과 Table 3.2와 같은 퇴원구간을 얻을 수 있다 (Park 등, 2010).

Table 3.2 Discharge intervals of explanatory variables

Variable name	Discharge interval
WBC	(4.5, 10.5)
PLT	(200, 520)
Cl-	(104, 110)
AST	(12, 48)
ALT	(0, 45)
$PCO2$	(26, 40)
$PO2$	(57, 96)
$O2SAT$	(94.4, 98.8)
LDH	(0, 300) $\cup$ (800,∞)
$Ca2+$	(2.1, 2.32)
$Mg2+$	(1.8, 2.3)

3.2. 벌점 로지스틱 회귀모형 적합 과정

3.1절에서 구한 이항 설명변수인 퇴원구간을 이용하여 호흡곤란 환자의 퇴원여부 결정을 위한 벌점 로지스틱 회귀모형을 적합하는 과정을 간략히 설명한다. 벌점 로지스틱 회귀모형을 적합시키기 위해 R의 glmnet 패키지를 사용하였다.

벌점 로지스틱 회귀모형을 적합시키는 첫 번째 단계는 식 (2.5)에 있는 최적의 λ 값의 결정이다. 이것을 위해 cv.glmnet 함수를 사용할 수 있는데, 이 연구에서는 10-묶음 교차타당성 (10-fold cross-validation)에 의해 최소의 이탈도 (deviance)를 가지는 λ를 이용하였다. 능형 로지스틱 회귀모형과 Lasso 로지스틱 회귀모형의 최적 λ 선정 과정을 보여주는 그림이 각각 Figure 3.1과 Figure 3.2에 주어져 있다.

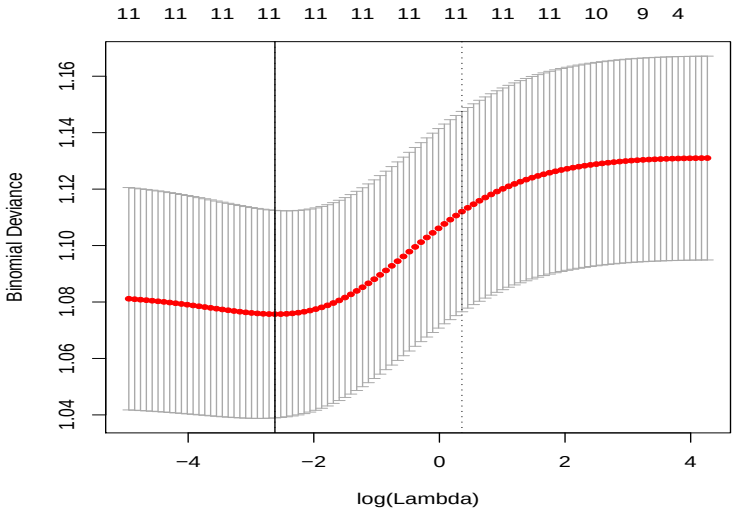

Figure 3.1 Optimal λ for the ridge logistic regression

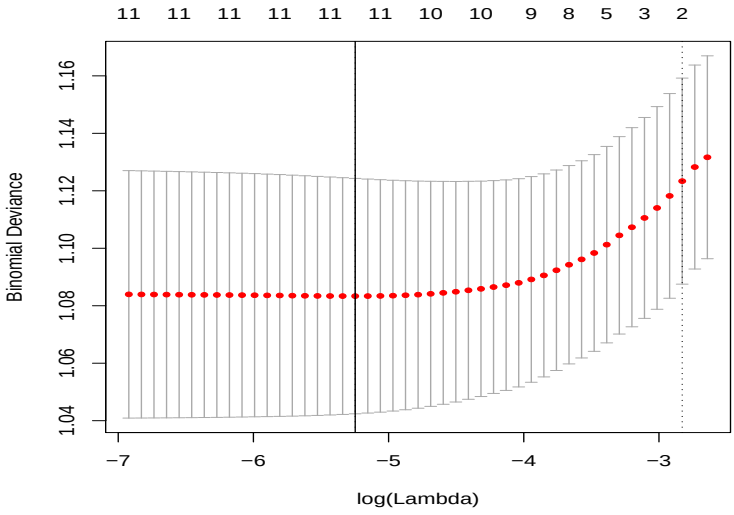

Figure 3.2 Optimal λ for the Lasso logistic regression

이 그림에는 횡축에 $\log(\lambda)$, 종축에 이탈도가 표시되어 있다. 각 그림에는 이탈도가 가장 작은 $\log(\lambda)$ 위치를 실선으로 표시하여 최적의 $\log(\lambda)$ 값이 대략 얼마인지 알아보기 쉽게 하였다. 이 그림에서 최적 $\log(\lambda)$가 능형 모형에서는 대략 -2.62, Lasso 모형에서는 대략 -5.25로 주어진 것을 알 수 있다.

이렇게 추정된 최적의 $\log(\lambda)$에 대해 식 (2.5)를 최소화시키는 회귀계수 값들을 구하게 된다. 능형 및 Lasso 모형에 대해 $\log(\lambda)$의 값에 따른 회귀계수 추정값들을 보여주는 그림은 각각 Figure 3.3과 Figure 3.4에 주어져 있다.

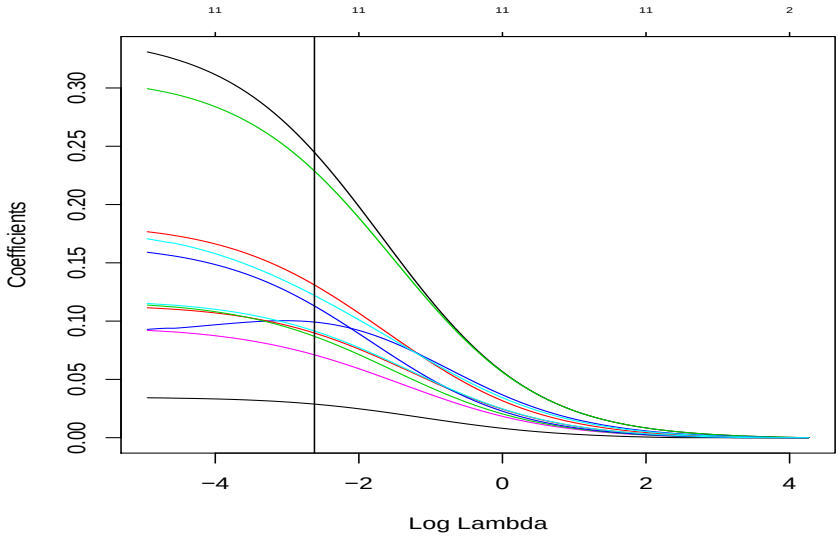

Figure 3.3 Regression coefficients for the ridge logistic regression

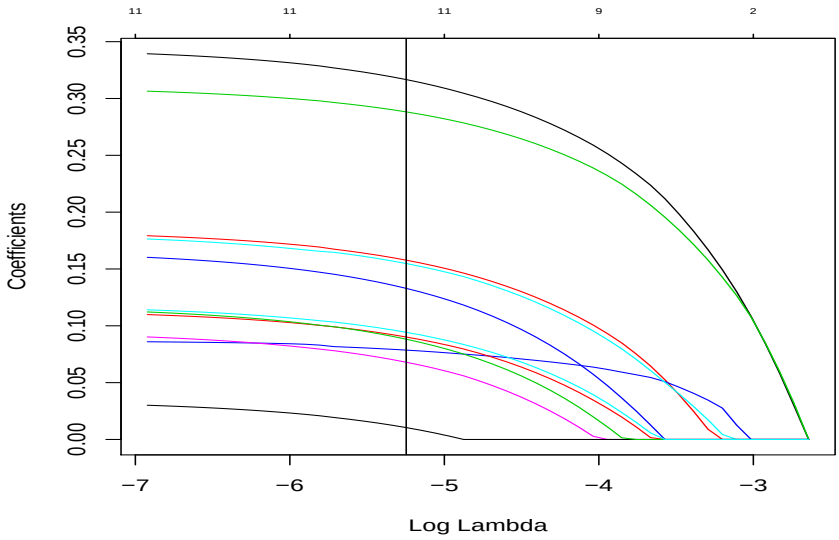

Figure 3.4 Regression coefficients for the Lasso logistic regression

이 그림에도 Figure 3.1과 Figure 3.2에서 구한 최적의 log(λ) 위치에 실선으로 표시하여 추정된 회귀계수 값을 알아보기 쉽게 하였다. 따라서 능형과 Lasso 모형 모두에서 회귀계수가 0으로 적합된 설명변수가 없다는 것을 알 수 있다. 따라서 해석력에서 어느 정도 강점이 있다고 알려진 Lasso 모형이 능형 모형보다 해석력이 나아진 것이 없음을 알 수 있다.

변수선택을 위해서는 R의 MMIX 패키지의 stepSel 함수를 이용하였다. 구체적으로 AIC (Akaike information criterion)를 기준으로 하고 단계적 방법으로 최적의 모형을 찾았더니 WBC, PLT, $Cl-$, ALT 및 $Ca2+$가 선택되었다.

전체 11개 설명변수가 들어간 모형 (All이라 표시됨), 변수선택된 모형 (Selected라고 표시됨), Lasso 모형 (Lasso라고 표시됨), 능형 모형 (Ridge라고 표시됨)의 회귀계수 추정값을 구하였더니 Table 3.3과 같았다.

Table 3.3 Regression coefficients for four logistic regression models

Variable name	All	Selected	Lasso	Ridge
Intercept	-1.209	-1.194	-1.183	-1.158
WBC	0.345	0.363	0.317	0.245
PLT	0.184	0.211	0.158	0.131
Cl-	0.311	0.337	0.288	0.229
AST	0.087	0.000	0.079	0.099
ALT	0.182	0.250	0.155	0.122
$PCO2$	0.096	0.000	0.068	0.071
$PO2$	0.035	0.000	0.010	0.029
$O2SAT$	0.115	0.000	0.090	0.090
LDH	0.118	0.000	0.088	0.087
$Ca2+$	0.167	0.165	0.133	0.113
$Mg2+$	0.119	0.000	0.094	0.092

Figure 3.3과 Figure 3.4에서는 시각적으로 능형 모형의 회귀계수 축소가 더 많이 일어난 것처럼 보였지만, Table 3.4에서 이것이 사실이 아니라는 것을 확인할 수 있다. 실제로 WBC, PLT, $Cl-$, ALT, $Ca2+$ 등에서는 능형 모형의 회귀계수 축소가 더욱 크게 일어났으나, 나머지 변수에서는 Lasso 모형의 회귀계수 축소가 조금 더 크게 일어나거나 혹은 두 모형의 축소 차이가 미미하게 나타났다. 따라서 이 자료에서는 변수선택에 의해 선택된 변수에 대해서 능형 모형의 회귀계수가 더욱 크게 축소되는 현상을 관측할 수 있었다.

마지막으로 네 가지 모형의 예측정확도를 비교하기 위해서 10-묶음 교차타당성을 이용하여 이탈도를 비교하였는데 그 결과는 Table 3.4에 주어져 있다.

Table 3.4 Prediction accuracy based on deviance for four logistic regression models

Models	All	Selected	Lasso	Ridge
Deviance	721.75	732.61	722.28	718.49

이 표에 의하면 Ridge > All ≈ Lasso >> Selected 순서로 예측정확도가 뛰어난 것을 알 수 있다. 혈액검사 결과 간의 연관성이 존재할 것이기 때문에 능형 모형의 예측정확도가 높은 것은 어느 정도 예상해 볼 수 있었지만, 모든 변수가 다 들어간 모형의 예측정확도가 Lasso 모형이나 변수선택된 모형보다 높게 나타난 것은 예상하지 못했던 일이다. 이 현상은 3.1절에서 퇴원구간을 만드는 과정에서 반응변수와의 관련성이 고려된 이산형 설명변수인 퇴원구간이 생성되면서 개개 설명변수의 예측력이 높아진 데 기인한 것이 아닐까 생각한다.

4. 결론 및 논의

호흡곤란의 원인질환은 짧은 시간의 문진으로 진단을 하기 어렵다. Park 등 (2010)에서는 호흡곤란 환자를 대상으로 11개 혈액검사 결과에 기반한 퇴원 결정을 위한 간편한 통계모형을 제안하였다. 또한 Park (2011)에서는 수량화 기법을 이용하여 11개 혈액검사의 중요성을 고려한 통계모형을 제안하였다. 이 방법은 결국 Fisher의 선형판별함수 (linear discriminant function)에 근거한 통계모형이 된다.

이 연구는 혈액검사 결과에 근거하여 호흡곤란 환자의 퇴원여부 예측력을 더욱 높이려는 시도에서 시작되었는데 벌점 (이항) 로지스틱 회귀모형이 이용되었다. 구체적으로 이 연구에서 사용된 것은 Lasso와 능형이 포함된 벌점 로지스틱 회귀모형이었다. 벌점 로지스틱 회귀모형과 비교대상이 된 모형으로는 일반 로지스틱 회귀의 11개 혈액검사 전체를 사용한 모형과 변수선택에 의한 모형이 있다. 그리고 설명변수의 예측력을 높이기 위해서 Park 등 (2010)에 의해 설정된 (혈액검사 결과들의) 퇴원구간에 근거한 이분형 변수를 사용하였다.

벌점 로지스틱 회귀모형의 적합과정은 크게 두 단계로 나뉜다. 첫 번째 단계에서는 (2.5)의 벌점을 결정해주는 λ의 최적값을 구하는 과정이다. 두 번째 단계는 이렇게 구한 λ 값에 해당되는 회귀계수 값을 구하는 과정이다.

벌점 로지스틱 회귀모형에서 구한 회귀계수 추정값을 일반 로지스틱 회귀의 모든 변수가 들어간 모형과 변수선택된 모형과 비교하였더니 다음과 같은 결과를 얻었다. 벌점 로지스틱 회귀모형의 회귀계수 추정값은 모두 0이 아니어서 해석력 관점에서 전체변수를 사용한 모형보다 개선된 것이 없었다. 그러나 10-묶음 교차타당성에 의한 예측정확도 관점에서는 능형 모형, 전체변수 모형 혹은 Lasso 모형, 변수선택 모형 순서로 나타났다.

능형 모형과 Lasso 모형의 회귀계수 축소에는 다음과 같은 경향이 있었다. 능형 모형에서 변수선택에 의해 선택된 변수의 회귀계수가 더 많이 축소되었으며, 그 외의 변수의 회귀계수는 Lasso 모형에서 다소 많이 축소되거나 혹은 두 모형의 회귀계수 축소가 비슷하게 일어났다. 이것이 호흡곤란 환자 자료에 국한되어 발생하는 현상인지 아니면 보다 일반화할 수 있는 현상인지는 알 수 없지만, 더 많은 자료분석을 통해 관측해 볼 필요가 있다고 생각된다.

참고문헌

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous attributes as preprocessing for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022-1027.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1-22.

Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*, 3rd Ed., Prentice Hall, New Jersey.

Kerber, R. (1992). ChiMerge: Discretization of numeric attribute. *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92)*, 123-127.

Kim, J. S., Jang, Y. M. and Na, J. H. (2005) Comparison of multiway discretization algorithms for data mining. *Journal of the Korean Data & Information Science Society,* **16**, 801-813.

Lee, S., Park, J. E. and Oh, K. W. (2003) Discretization of continuous-valued attributes considering data distribution. *Journal of Korea Fuzzy Logic and Intelligent Systems Society,* **13**, 391-396.

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

Na, J. H., Kim, J. M. and Cho, W. S. (2005). Comparison of binary discretization algorithms for data mining. *Journal of the Korean Data & Information Science Society*, **16**, 769-780.

Park, C. (2011). A quantification study of blood test results for dyspnea patients. *Journal of the Korean Data & Information Science Society*, **22**, 477-485.

Park, C., Kim, T. Y., Kwon, O. J. and Park, H. S. (2010). A simple statistical model for determining the admission or discharge of dyspnea patients. *Journal of the Korean Data & Information Science Society*, **21**, 279-289.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **21**, 279-289.

Whitten, D. A. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society B*, **73**, 753-772.

Journal of the Korean Data & Information Science Society
2013, **24**(1), 125–133

http://dx.doi.org/10.7465/jkdi.2013.24.1.125
한국데이터정보과학회지

Penalized logistic regression models for determining the discharge of dyspnea patients

Cheolyong Park[1] · Myo Jin Kye[2]

[12]Department of Statistics, Keimyung University

Received 14 December 2012, revised 5 January 2013, accepted 10 January 2013

Abstract

In this paper, penalized binary logistic regression models are employed as statistical models for determining the discharge of 668 patients with a chief complaint of dyspnea based on 11 blood tests results. Specifically, the ridge model based on L^2 penalty and the Lasso model based on L^1 penalty are considered in this paper. In the comparison of prediction accuracy, our models are compared with the logistic regression models with all 11 explanatory variables and the selected variables by variable selection method. The results show that the prediction accuracy of the ridge logistic regression model is the best among 4 models based on 10-fold cross-validation.

Keywords: Cross-validation, determining discharge, dyspnea patients, penalized logistic regression.

[1] Corresponding author: Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea. E-mail: cypark1@kmu.ac.kr
[2] Master candidate, Department of Statistics, Keimyung University, Daegu 704-701, Korea.