LASSO를 이용한
간경변 발생 예측 모형 연구

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 성 은

LASSO를 이용한
간경변 발생 예측 모형 연구

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2006년 6월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 성 은

김성은의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2006년 6월 일

감사의 글

조금은 낯설기도 했던 대학원 생활이지만 무사히 마칠 수 있도록 도움을 주신 분들께 감사의 글을 전하고자 합니다.

먼저 2년 동안 부족한 저를 지도해주신 김동기 교수님께 감사드립니다. 의학통계에 발을 디딜 수 있게 도움을 주시고, 논문을 마칠 수 있도록 이끌어 주신 조진남 교수님과 논문 주제에 대한 기초를 가르쳐 주신 한광협 교수님께도 감사의 마음을 전합니다. 유전통계에 대한 이론을 가르쳐주신 임길섭 박사님, 배움에 대한 열정을 가르쳐주시고 바쁘신 와중에도 묵묵히 도움을 주신 김동건 교수님께 진심으로 감사드립니다.

항상 먼저 관심 가져 주시고 논문을 마무리할 수 있도록 도움을 주신 송기준 박사님과 배움에 있어서 한 번 더 생각하는 법을 가르쳐주신 명성민 박사님께 감사의 마음을 전합니다. 또한, 무영오빠의 위로와 격려가 많은 힘이 되었습니다. 자신의 일처럼 챙겨주신 미영언니에게는 고마운 마음과 함께 미안한 마음이 앞섭니다. 본받고 싶은 점이 많았던 찬미언니와 편하게 대해준 원열오빠, 자주 뵙지 못해도 먼저 챙겨주시는 명희언니, 작은 부분까지 신경써준 신영언니와 항상 밝은 미소로 즐겁게 해 준 수옥언니, 대학원 생활의 시작을 도와준 은정언니와 정숙언니에게 고마운 마음을 전합니다. 연구실 생활을 시작하고 함께 한 시간동안 항상 친절하게 대해주고 가르쳐준 혜리언니, 선배지만 동기 이상으로 편하게 대해준 소연이과 민진이 덕분에 외로운 대학원 생활을 무사히 마칠 수 있어서 항상 고마운 마음입니다. 후배지만 때로는 오히려 도움을 받은 은희씨, 친구처럼 편했던 혜진이에게 고마운 마음을 전하며 마무리도 잘 해낼거라 믿습니다. 많이 챙겨주지 못해 미안한 정윤씨와 수희, 진희, 경화, 영애에게 남은 대학원 생활이 후회 없는 시간이 되길 바랍니다.

항상 곁에서 서로를 먼저 생각해주는 친구 은정, 경진, 지숙에게 고맙고, 앞으로도 서로에게 웃음을 줄 수 있는 친구로 남길 바랍니다. 대학원 생활이 힘들 때마다 많은 위로를 해 주었던 소영, 지수, 일호, 영미에게도 고맙다는 말을 전하고

싶습니다. 항상 먼저 연락하지 못해 미안한 수진과 진미, 무뚝뚝하지만 재미있는 연태, 바쁘게 살고 있는 태용, 자주 보지는 못해도 마음만으로도 든든한 정화, 효성, 관석, 웅용, 영숙, 경옥, 그리고 항상 만나면 즐거운, 오랜 친구 영주가 있어서 고맙고 든든합니다.

항상 웃음을 주어 즐겁게 해주는 큰 언니와 작은 언니가 있어서 대학 생활과 대학원 생활을 즐겁게 보낼 수 있었습니다. 앞으로는 언니들에게도 좋은 일만 생기길 바라며, 이제 곧 사회인이 되는 막내 동생에게도 좋은 결실이 맺어졌으면 좋겠습니다. 마지막으로 제가 원하는 것을 할 수 있도록 항상 저를 믿어 주시고 후원해 주시는 부모님께 진심으로 감사드립니다. 그리고 사랑합니다.

2006년 6월

김 성 은 올림

차 례

표 차 례

그 림 차 례

국문요약

LASSO를 이용한 간경변 발생 예측 모형 연구

간경변증(Liver Cirrhosis)은 간의 염증이 오래 지속되어 간의 정상적인 구조가 소실되고 굳어지면서 간의 표면이 울퉁불퉁해지는 것으로, 많은 합병증을 유발시킬 수 있는 질병이므로 건강검진 등의 조기진단을 통해 간경변 발생을 줄이기 위한 노력이 필요하다.

본 논문에서는 1994년 5월 ~ 2005년 9월에 건강검진을 받은 85,458명의 검진자 중 병원에 내원하여 소화기내과 검사를 받은 4,093명을 대상으로 간경변 발생에 대한 위험인자를 살펴보고, 간경변 발생 예측 모형을 연구하였다. 로지스틱 회귀모형에 LASSO를 적용하여(LASSO-logistic regression) 예측모형으로 사용하였으며, 기존의 데이터마이닝 기법인 판별분석, 단계적 로지스틱 회귀분석, 의사결정나무, Random Forests, SVM-RFE와 성능을 비교하였다. 그 결과, LASSO-logistic regression을 통해 구축된 간경변 발생 예측 모형의 성능은 정확도가 91.6%, 민감도가 58.27%로 나타나 기존의 데이터마이닝 기법과 비슷한 성능을 보였다. 간경변 발생 위험인자는 B형간염S항원(HBsAg), C형간염항체(AntiHCV), 가족력, 음주력, 혈소판 수치(Platelet), 알칼라인 포스파타제(Alk.Phos), 알부민(Albumin), $\gamma-GT$로, 기존에 알려진 간경변 발생 위험인자를 잘 나타내주고 있으며, 의사결정나무, Random Forests, SVM-RFE와 달리 변수의 해석이 쉽다. 그러므로 LASSO-logisitic regression은 변수선택과 판별분석이 동시에 필요한 자료에 적합하다고 할 수 있다.

핵심 되는 말 : 간경변, 건강검진, 위험인자, 예측모형, 민감도, 정확도, 축소, 변수선택, LASSO

1장 서 론

1.1 연구배경

간경변증(Liver Cirrhosis)은 만성간염으로 간의 염증이 오래 지속된 결과 두꺼운 섬유질이 형성되고, 살아남은 간세포들에 의해 재생결절이 형성되면서 간의 정상적인 구조가 소실되고 굳어지면서 간의 표면이 울퉁불퉁해지는 것을 말한다.
간경변증은 있으나 합병증을 동반하지 않고 임상적으로 괜찮은 상태인 대상성 간경변증이 있는 반면, 각종 합병증을 동반하는 비대상성 간경변증이 있다. 간경변증의 합병증에는 복수, 출혈, 노폐물 축적으로 인한 혼수, 자발성 복막염 등이 있다. 따라서 대상성 간경변증인 경우에는 큰 문제가 되지 않지만 비대상성 간경변증이라면 각종 합병증이 나타나기 쉽고, 합병증 자체로 앓아눕거나 사망할 가능성이 적지 않으므로 세심한 주의가 필요하다.
그러나 많은 환자들은 간경변이 한참 진행될 때까지 별다른 증상을 느끼지 못한다. 간은 15~20%만 있어도 최소한 생존에 필요한 대사 작용을 해내기 때문이다. 이 같은 간의 특성 때문에 오히려 진단이 늦어지고 일단 진단이 내려져도 환자들이 관리를 소홀히 하고 있다. 그러므로 간경변 발생에 대한 조기진단을 통해 간경변 발생을 줄이기 위한 노력이 요구된다.

1.2 연구목적 및 내용

본 논문에서는 1994년부터 2005년까지 건강검진센터에서 건강검진을 받은 85,458명의 검진자 중 병원에 내원하여 소화기내과 검사를 받은 4,093명을 연구대상으로 하였다. 간경변 진단을 받은 검진자의 건강검진 자료(screening test data)를 토대로 간경변 발생을 예측할 수 있는 인자를 살펴보고 간경변 발생 예측을 위한 통계학적 모형을 구축하였다.

예측 모형 추정에는 변수선택과 판별을 동시에 해주는 최소절대축소선택연산(Least Absolute Shrinkage and Selection Operator, 이하 LASSO)를 사용하였고, 기존의 데이터마이닝(datamining) 기법인 판별분석(Discriminant Analysis), 단계적 로지스틱 회귀분석(Stepwise Logistic Regression), 의사결정나무(Desicion Tree), Random Forests, SVM-RFE와 그 성능을 비교 분석하였다.

본 논문의 구성은 **1**장 서론에서는 연구 배경, 목적 및 내용에 대해 제시하였고, 2장에서는 기존의 데이터마이닝 기법들을 소개하였다. 3장에서는 LASSO의 이론적인 내용을 기술하였고, 4장에서는 건강검진 자료를 소개하였다. 5장에서는 2장과 3장에서 소개한 기법들을 건강검진 자료를 통해 비교하였으며, 6장에서 결론 및 고찰에 대해서 논의하였다.

2장 여러 가지 판별 모형

2.1 판별분석

판별분석(discriminant analysis)은 이미 알려진 두 개 이상의 집단에 속하는 관찰값으로부터 각 집단의 차이를 잘 설명하여 줄 수 있는 독립변수들의 선형결합을 찾고, 이 함수식에 따라 소속집단이 알려지지 않은 새로운 개체를 분류(classification)하는 기법이다.

관찰값들이 다변량 정규분포를 따르는 경우, 판별분석은 두 집단의 공분산행렬의 동일성 여부에 따라 두 가지로 나뉠 수 있다. 두 집단의 공분산행렬이 같은 경우에는 선형판별분석(Linear Discriminant Analysis, 이하 LDA)이고, 두 집단의 공분산행렬이 다른 경우에는 이차판별분석(Quadratic Discriminant Analysis, 이하 QDA)이 된다.

(**1**) 선형판별분석

두 집단의 확률밀도함수 $f_i(x)$는 평균벡터가 μ_i이고 공분산행렬이 Σ_i인 다변량 정규밀도함수라고 가정하자.
두 공분산행렬이 같은 경우 두 확률밀도함수의 비율에 자연대수를 취하면 다음과 같이 된다.

$$\begin{aligned} L(x) &= \ln\left\{\frac{f_1(x)}{f_2(x)}\right\} = \ln f_1(x) - \ln f_2(x) \\ &= -\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) \\ &= (\mu_1-\mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \end{aligned}$$

모집단의 모수들이 일반적으로 알려져 있지 않기 때문에 μ_1, μ_2, Σ는 표본으로부

터 추정해야 한다. 모평균 μ_i의 추정치인 표본평균벡터 $\overline{x_i}$와 모분산 Σ의 불편추정치인 합동표본공분산행렬을 S_p라 할 때 판별함수는 다음과 같다.

$$\hat{L}(x) = (\overline{x_1} - \overline{x_2})^T S_p^{-1} x - \frac{1}{2}(\overline{x_1} - \overline{x_2})^T S_p^{-1}(\overline{x_1} + \overline{x_2})$$

위의 판별함수는 관측치 x에 대한 선형식으로 표시되며, 이것을 표본선형분류함수 혹은 표본선형판별함수라고 부른다.
각 집단의 사전확률을 π_i라고 하면 다음과 같은 조건일 때 새로운 관측치 x_0는 집단 g_1에 분류하고, 그 이외의 경우에는 집단 g_2에 분류하게 된다.

$$\hat{L}(x_0) \geq \ln\frac{\pi_2}{\pi_1}$$

(2) 이차판별분석

두 집단의 공분산행렬이 다른 경우 판별함수는 공분산행렬이 같은 경우보다 약간 복잡한 함수형태가 된다.

$$\begin{aligned} Q(x) &= \ln\left\{\frac{f_1(x)}{f_2(x)}\right\} = \ln f_1(x) - \ln f_2(x) \\ &= -\frac{1}{2} ln(|\Sigma_1|/|\Sigma_2|) - \frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) \\ &= -\frac{1}{2}x'(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x - k \end{aligned}$$

두 집단의 모수가 알려져 있지 않은 경우 모평균벡터 μ_i는 표본평균벡터 $\overline{x_i}$로, 공분산행렬 Σ_i는 표본공분산행렬 S_i로 추정되며 판별함수는 다음과 같이 나타낼 수 있다.

$$\hat{Q}(x) = -\frac{1}{2}x^T(S_1^{-1} - S_2^{-1})x + (\overline{x_1}S_1^{-1} - \overline{x_2}S_2^{-1})x - k$$

여기서 $k = \frac{1}{2}ln(|S_1|/|S_2|) + \frac{1}{2}(\overline{x_1}S_1^{-1}\overline{x_1} - \overline{x_2}S_2^{-1}\overline{x_2})$을 나타낸다.

각 집단의 사전확률을 π_i라고 하면 다음과 같은 조건일 때 새로운 관측치 x_0는 집단 g_1에 분류하고, 그 이외의 경우에는 집단 g_2에 분류하게 된다.

$$\hat{Q}(x_0) \geq \ln\frac{\pi_2}{\pi_1}$$

2.2 단계적 로지스틱 회귀분석

로지스틱 회귀분석(logistic regression)은 종속변수가 두 가지 값만 취하는 질적인 이분형(binary) 변수일 때, 확률에 대해 로짓변환(logit transformation)을 하여 분석하는 방법이다.

p개의 독립변수에 대해서 종속변수가 **1**을 가질 확률을 $P(Y=1|x_1, x_2, ..., x_p)$라고 할 때, 로짓변환(logit tranasformation)은 다음과 같다.

$$\log\frac{P(Y=1|x_1,...,x_p)}{1-P(Y=1|x_1,...,x_p)} = \beta_0 + \beta_1 x_1 +, ...,+ \beta_p x_p$$

이것을 로지스틱 반응함수라고 부르고 다음과 같이 정리할 수 있다.

$$P(Y=1|x_1, x_2, ..., x_p) = \frac{exp(\beta_0 + \beta_1 x_1 +, ...,+ \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 +, ...,+ \beta_p x_p)}$$

로지스틱 반응함수의 계수들을 추정하기 위해 최대우도 추정(maximum likelihood estimation)이 일반적으로 많이 쓰이며 이것은 모집단으로부터 표본 관측값이 주어졌을 때 로그우도함수의 최대화로부터 얻어진다.

회귀계수의 최대우도 추정값은 비선형이기 때문에 피셔의 스코어링 방법(Fisher's method of scoring)이나 뉴튼-랩슨 방법(Newton-Raphson method)과 같은 반복적인 방법(iterative method)에 의해 해의 근사값을 구해야한다.

회귀계수의 최대우도 추정값이 구해지면, 각 개체에 대한 사후확률(posterior probability), $\widehat{p_x}$을 이용하여 분류를 할 수 있다. 0과 **1** 사이의 값을 갖는 추정된 사후확률에 대해 적절한 경계값(cut off value), c를 정하여 이 값을 기준으로 다음과 같이 분류한다.

$$\begin{cases} \widehat{p_x} \geq c \text{ 이면, 집단 1로 분류} \\ \widehat{p_x} < c \text{ 이면, 집단 0으로 분류} \end{cases}$$

입력변수의 수가 많은 경우에는 로지스틱 회귀분석에서 변수를 선택해야한다. 변수를 선택하는 방법에는 입력변수를 각 변수의 기여도에 따라서 하나씩 추가하면서 선택하는 전진 선택법(forward selection)과 모든 변수를 포함하는 완전모형으로부터 시작하여 불필요한 변수를 하나씩 제거해 나가는 후진 제거법(backward elimination)이 있다. 전진 선택법은 한 번 선택된 변수는 모형에서 제거될 수 없고, 후진 제거법은 한 번 제거된 변수는 모형에 다시 선택될 수 없는 단점이 있다. 반면에 단계적 선택법(stepwise selection)은 전진선택법에 후진 제거법을 결합한 것으로서, 매 단계마다 선택과 제거를 반복하면서 중요한 변수를 찾는 방법이다. 일반적으로 가장 널리 쓰이며, 본 논문에서도 단계적 선택법을 적용하여 단계적 로지스틱 회귀분석(Stepwise Logistic Regression)을 시행하였다.

2.3 의사결정나무

의사결정나무(decision tree) 모형은 표본 집단을 특정 기준값에 의해 유사한 집단으로 분류하고, 분류된 하위집단을 다시 특정 기준을 찾아 분류하는 과정을 반복함으로써 종속변수와 독립변수들 간의 패턴이나 관계를 찾아내는 분석방법이다. 종속변수에 가장 큰 영향을 주는 독립변수의 특정값을 기준으로 표본 집단에 대한 최초 분리가 이루어지며 순차적(sequential)으로 더 이상의 분리가 이루어지지 않을 때까지 분리를 수행한다.

의사결정나무 모형은 다양한 알고리즘에 의해 분리가 이루어지고, 이런 과정은 나무구조로 표현되며, 나무구조가 시작되는 뿌리마디(root node), 하나의 마디로부터 분리되어 나간 두 개 이상의 마디들인 자식마디(child node), 자식마디의 상위마디인 부모마디(parent node), 각 나무줄기의 끝에 위치하고 있는 끝마디(terminal node) 등 여러 가지의 마디(node)라고 불리는 구성요소들로 이루어져 있다(Quinlan J R. 1986).

분리기준(split criterion)에 의해 부모마디로부터 자식마디들이 형성되는데, CHAID, CART, C4.5 등의 알고리즘에 의해서 분리기준이 정해진다. 이 논문에서는 마디(node)의 순수함(purity)을 나타내는 지니계수(Gini index)에 의해 분리여부를 결정하는 CART(Classification And Regression Tree) 알고리즘을 사용하였다. 지니계수는 집단이 순수할수록 값이 작아지게 된다.

의사결정나무 모형은 분류 또는 예측의 과정이 나무구조에 의해서 표현되기 때문에 해석이 편리하다. 또한 변수들 간의 상호작용효과(interaction effects)를 파악할 수 있다는 장점도 가지고 있다. 그러나 연속형 변수(continuous variable)를 비연속적인 값으로 취급하기 때문에 분리의 경계점 부근에서는 오류가 발생할 확률이 높으며, 주효과(main effect)를 얻을 수 없다는 단점이 있다. 독립변수의 주효과를 밝혀낼 수 있는 판별분석이나 회귀분석의 경우와는 달리 의사결정나무는 변수간의 상호작용효과만을 보기 때문에 한계가 있다.

2.4 Random Forests

Random Forests는 브레이만(Leo Breiman, 2001)에 의해서 개발된 분류 알고리즘으로서 다양한 붓스트랩 표본으로부터 얻은 나무 분류자들을 하나의 분류자로 만드는 기법이다. Random Forests는 단일 나무를 만드는 대신 여러 개의 나무로 확장시킴으로써 분류의 정확도를 높이고자 한 것이다.

Random Forests는 다음의 두 가지 random 성분을 모두 포함하는 알고리즘이다. 첫째, 각 나무들은 학습용 자료에서 무작위로 뽑은 다양한 붓스트랩 표본으로부터 나무를 만들고 둘째, 나무를 만들 때 변수 모두를 사용하는 것이 아니라 무작위로 선택하여 모형을 구축하는 무작위 변수 선택이다.

$h_1(\underline{x}), ..., h_k(\underline{x})$를 분류자의 집단, X와 Y를 분석용 자료(training set)에서 추출된 확률벡터(random vector)라고 할 때 margin function은 다음과 같이 정의된다.

$$mg(X, Y) = av_k I(h_k(X) = Y) max_{j \neq Y} av_k I(h_k(X) = j)$$

Margin function은 분류자가 Y를 옳게 분류한 사건의 수의 평균값과 오분류한 사건의 평균값의 초과량을 측정한다. 따라서 margin이 크면 클수록 분류에 대해서 더 많이 신뢰할 수 있다.

Random Forests에서는 변수의 중요도(importance)를 판단할 수 있는데, 변수가 모형에 추가되었을 때 지니계수(Gini index) 감소율의 정도로 중요 변수를 평가하거나 변수가 모형에서 제거되었을 때 정확도가 감소하는 정도로 평가한다.

2.5 SVM-RFE

Vapnik의 Support Vector Machine(SVM)은 선형판별함수(Linear discriminant function)의 개념을 커널함수를 이용하여 비선형화 한 모형이다. p 차원의 입력변

수 $x \in R^p$를 이용하여 두 개의 수준을 가지는 목표변수 $y \in \{-1, 1\}$를 예측하는 판별모형을 상정할 때, 선형판별함수는 다음과 같이 정의된다(Hastie et al. 2002).

$$f(x) = sign(w^T x + b)$$

여기서 $sign(.)$은 부호함수, $w \in R^P$는 가중치벡터, b는 절편을 나타낸다. margin은 $2/\|w\|$로 정의되며, 각 그룹의 데이터로 convex hull을 구성했을 때 convex hull 간의 최단거리를 말한다. margin은 가중치벡터 w의 길이에 반비례하는데, 선형 SVM은 이 margin을 최대화시키는 w를 찾는 방법으로, 이차 프로그래밍(Quadratic Programming)을 이용하여 SVM의 해를 구할 수 있다.

그러나 SVM은 변수선택을 할 수 없기 때문에 Guyon 등이 제안한 RFE(Recursive Feature Elimination)를 이용하여 변수 선택을 할 수 있다. SVM에서 판별은 $w^T x + b$의 부호로 결정되므로 $|w_j|$이 클수록 x_j가 $f(x)$에 미치는 영향은 커지게 된다. RFE는 이것을 이용하여 w_j에 대해 $c_j = w_j^2$으로 정의하고 SVM에 적용하여 입력변수의 중요도를 구한다. RFE는 변수의 중요도를 쉽게 구할 수 있으나 일반적인 후진 제거법의 단점을 그대로 가지게 된다. 즉, 한 번 제거된 변수는 다시 모형에 포함될 수 없으므로, 입력 변수들 간의 교호작용이 있을 경우 그 관계를 파악하는 것이 어렵다.

3장 LASSO

3.1 이론적 배경

다음과 같은 회귀 모형이 있다고 하자.

$$y_i = \alpha + x_i^T \beta + \epsilon, \quad i = 1,2,\dots,p$$

일반적으로 이 문제를 해결하기 위해 잔차의 제곱합을 최소화하는 보통최소제곱(Ordinary Least Squares, OLS)이 사용된다. 그러나 보통최소제곱은 독립변수의 개수가 증가하면, 독립변수들 사이의 강한 상관관계(correlation)로 인한 다중공선성(multicollinearity)이 존재할 수 있으며, 따라서 회귀계수의 분산이 커져서 추정 회귀식의 예측력(prediction accuracy)이 떨어지는 문제점이 있다. 또한 독립변수의 개수가 증가하면 변수에 대한 해석력(interpretation)이 떨어진다. 즉, 많은 독립 변수 중 어떤 변수가 중요한 역할을 하는지에 대한 판단이 어려워진다.

보통최소제곱의 문제점이 보완된 방법으로 부분집합선택방법(subset selection method)과 능형회귀(ridge regression)가 있다. 부분집합선택방법(subset selection method)은 해석 가능한 모형을 제공하지만 데이터가 조금만 바뀌어도 전혀 다른 모형이 나타날 수 있고 이것은 예측력(prediction accuracy)을 떨어뜨릴 수 있다. 능형회귀(ridge regression)는 회귀 계수들에 제약조건, $\Sigma\beta_j^2 \le t$을 주어 계수를 추정하는 방법으로 회귀계수를 축소(shrinkage)시켜 분산이 작아지지만 회귀계수를 완전히 0으로 추정하지는 못하므로 모형의 해석이 쉽지 않다.

LASSO(Least Absolute Shrinkage and Selection Operator)는 Tibshirani(1996)에 의해 개발된 방법으로, 회귀 계수들에 제약조건, $\Sigma|\beta_j| \le t$을 주어 계수 추정치들의 크기를 축소(shrinkage)시키는 동시에 변수선택도 할 수 있는 추정방법이다.

3.2 모형의 설정

p개의 독립변수 $x_i = (x_{i1,\ldots,} x_{ip})^T$와 종속변수 y_i를 가지는 다음과 같은 모형이 있을 때,

$$y_i = \alpha + x_i^T\beta + \epsilon \quad i = 1,2,\ldots,p$$

회귀계수 추정치를 $\hat{\beta}= (\widehat{\beta_1}, \widehat{\beta_2},\ldots, \widehat{\beta_p})^T$이라고 하면, LASSO에 의한 회귀계수 추정 $(\hat{\alpha}\ ,\ \hat{\beta})$은 다음을 만족한다.

$$\arg\ min\ \{\sum_{i=1}^{N}(y_i - \alpha - \sum_j \beta_j x_{ij})^2\}$$

$$subject\ to\ \sum_j |\beta_j| \leq t$$

여기서 $t\,(\geqq 0)$는 회귀계수의 축소의 정도를 조절하는 조절모수(tuning parameter)이다. 독립변수 x_{ij}는 평균이 0, 분산이 **1**로 표준화(standardization)되었다고 가정하면, 모든 t에 대해서 다음을 만족한다.

$$\begin{aligned}\frac{\partial\left[\sum_{i=1}^{N}(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij})^2\right]}{\partial\alpha}\Big|_{\alpha=\hat{\alpha}} &= -2\sum_{i=1}^{N}(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij})\Big|_{\alpha=\hat{\alpha}} \\ &= -2\sum_{i=1}^{N}(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij}) \\ &= 0\end{aligned}$$

모든 j에 대해서 $\Sigma_{i=1}^{N} x_{ij} = 0$ 이므로, $\hat{\alpha}$은 다음과 같다.

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{j=1}^{p}\hat{\beta}_j\left(\sum_{i=1}^{N} x_{ij}\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N} y_i$$

일반화손실(loss of generality) 없이 $\bar{y} = 0$을 가정할 수 있으므로 α는 생략할 수 있다. 따라서 LASSO에 의한 회귀계수 추정은 다음을 만족하는 $\hat{\beta}$을 찾는 것이다.

$$\arg\, min \{ \sum_{i=1}^{N} (y_i - \sum_{j} \beta_j x_{ij})^2 \}$$

$$subject\ to\ \sum_{j} |\beta_j| \leq t$$

LASSO 추정 방법의 제약조건을 L**1**-제약조건(L**1**-constraint)이라고 부르며, 제약조건의 특성상 t가 줄어듦에 따라 중요하지 않은 변수들의 회귀계수들부터 차례로 0을 만들게 된다. 조절 상수 t는 그 값이 충분히 커질 때 계수에 대한 아무런 제약을 주지 않게 되어 일반적인 회귀 모형(regression model)과 같은 결과를 주게 된다.

그림**1-1**과 **1**-2는 LASSO가 회귀계수를 0으로 축소시키는 것을 보여준다. 그림**1-1**의 점선은 보통최소제곱에 의한 추정치를, 실선은 LASSO에 의한 추정치를 나타낸다. 그림**1**-2는 능형회귀의 제약조건과 비교한 것으로 왼쪽의 사각형은 LASSO의 제약조건 $|\beta_1| + |\beta_2| \leq t$을, 오른쪽의 원모양은 능형회귀의 제약조건 $\beta_1^2 + \beta_2^2 \leq t$을 의미한다. 등고선 중심에 있는 $\hat{\beta}$은 보통최소제곱추정치이고, 타원형등고선(elliptical contours)은 잔차제곱합(residual sum of squares)을 나타낸다. LASSO의 경우 사각형과 닿는 지점에서 추정된 회귀계수가 0이 될 수 있다. 그러나 능형회귀의 경우에는 추정된 회귀계수가 0이 될 수 없다는 것을 보여준다.

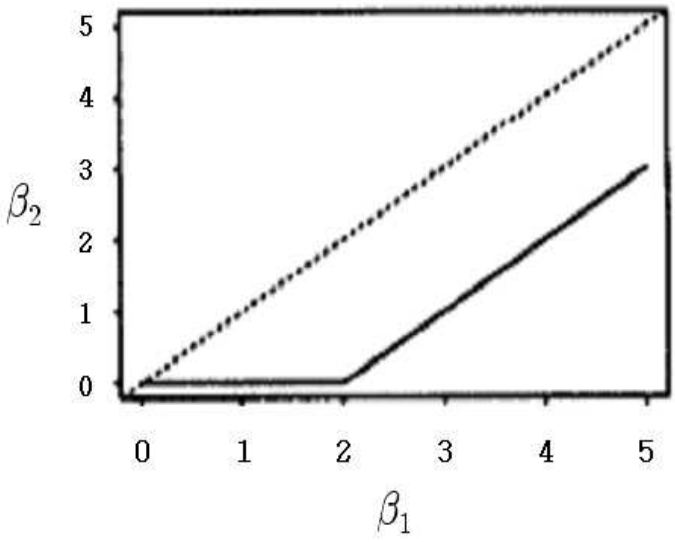

그림**1-1**. LASSO의 회귀계수

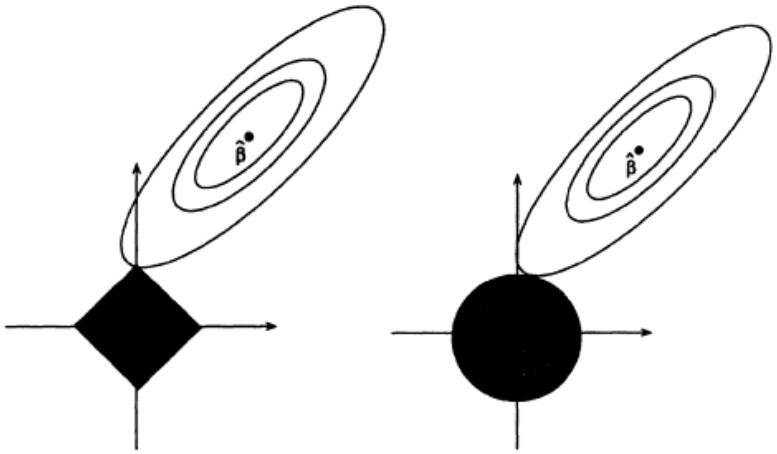

그림**1**-2. LASSO의 회귀계수

3.3 조절모수의 추정

LASSO에서 조절모수 t를 추정하기 위해 교차타당성(cross-validation, CV)이 사용된다.

교차타당성은 예측오차(prediction error)를 추정하는 데 널리 사용되는 방법이다. 자료를 임의의(random) k개로 나누어 $k-1$개의 자료는 훈련용 자료(training set)로 사용되고, k개의 자료 중 하나는 검증용 자료(test set)로 사용된다. 즉, $k-1$개의 훈련용 자료로 모형을 만들고 검증용 자료로 모형의 타당성을 검정하며, 이것을 k번 반복하여 평균제곱오차(Mean Squared Error, MSE)를 계산한다.

조절모수 t를 변화하면서 교차타당성을 실시한 후 가장 작은 평균제곱오차를 가지는 t를 선택하거나, **1**-Standard Error($1-SE$) 원칙을 적용하여 가장 작은 평균제곱오차의 표준오차(standard error, SE)만큼의 범위에 포함되는 t 중에서 가장 단순한 모형을 갖는 t를 선택한다.

3.4 회귀계수의 추정

LASSO는 회귀계수들의 절대값의 합이 조절모수 t보다 작거나 같도록 제약조건을 주어 회귀계수를 추정하게 된다. 그러나 이것은 미분 가능하지 않기 때문에 비선형 알고리즘(Nonlinear algorithm) 등을 이용하여 해를 구해야 한다. 이 논문에서는 LASSO의 회귀계수를 추정하기 위한 두 가지 알고리즘을 소개하기로 한다.

LASSO의 회귀추정은 β에 대해서 가능한 부호(sign)의 조합에 해당되는 2^p개의 부등식 제약조건(inequality constraints)를 가진 최소제곱에 관한 문제로 표현될 수 있다. 즉, 제약조건을 행렬(matrix)로 나타내면 $G\beta \le t$와 같이 나타낼 수 있는데, $G_{2^p \times p}$는 $(\pm 1, \pm 1, ..., \pm 1)$로 구성되어 있다. 이 부등식 제약조건을 다음과 같이 순차적으로(sequentially) 해결함으로써 LASSO의 해를 구할 수 있다(Lawson

and Hansen, 1974).

$g(\beta)=\sum_{i=1}^{N}(y_i-\Sigma_j\beta_j x_{ij})^2$, $i=1,2,\ldots,2^p$에 대해서 δ_i는 $(\pm1,\pm1,\ldots,\pm1)$로 구성되어 있다고 할 때, 제약조건 $\Sigma|\beta_j|\le t$은 $\delta_i^T\beta\le t$와 같게 된다. 이 제약조건을 다음과 같이 E와 S로 나타낸다.

$$E=\{i:\delta_i^T\beta=t\}\quad,\ S=\{i:\delta_i^T\beta<t\}$$

E는 회귀계수의 절대값의 합이 조절모수 t와 같은 등식제약식집합(equality set)이고, S는 회귀계수의 절대값의 합이 조절모수 t보다 작은 집합(slack set)을 의미한다.

$i\in E$에 대해서 G_E는 행(row)이 δ_i인 행렬을 나타내고, 1은 행렬 G_E의 행의 수와 같은, 1로 이루어진 벡터(vector)를 나타낼 때 LASSO의 회귀계수 추정을 위한 알고리즘을 다음과 같이 정리할 수 있다.

(**1**) $\delta_{i_0}=sign(\widehat{\beta^0})$인 $E=\{i_0\}$로 시작한다. $\widehat{\beta^0}$은 최소제곱추정치(overall least squares estimate)이다.

(2) $G_E\beta\le t1$을 제약조건으로 하는 $g(\beta)$를 최소화하는 $\hat{\beta}$을 찾는다.

(3) $\sum|\widehat{\beta_j}|\le t$ 을 만족하는 경우, 회귀계수 추정을 끝낸다.

$\sum|\widehat{\beta_j}|\le t$ 을 만족하지 않는 경우, 위배된 제약조건 i를 E에 추가한다.

(4) $G_E\beta\le t1$을 제약조건으로 하는 $g(\beta)$를 최소화하는 $\hat{\beta}$을 찾는다.

위의 과정은 2^p개의 제약조건을 가지므로 2^p번의 반복(iteration)이 일어난 후 수렴하게 된다.

두 번째 알고리즘은 David Gay에 의해 제안되었다. β_j를 $\beta_j^+-\beta_j^-$로 표현하는

방법이다. 즉, 제약조건 $\beta_j^+ \geq 0$, $\beta_j^- \geq 0$와 $\sum_{j=1}^{p}(\beta_j^+ + \beta_j^-) \leq t$를 가진 최소제곱이 되어 첫 번째 알고리즘에서의 변수 p개, 제약조건 2^p개를 가진 추정이 변수 $2p$개, 제약조건 $2p+1$개를 가진 추정으로 바뀌게 된다. 따라서 $2p+1$번의 반복이 일어난 후 수렴하게 되며, 회귀계수를 추정하기 위한 과정은 첫 번째 알고리즘과 같다.

3.5 LASSO의 로지스틱 회귀모형

종속변수가 이분형 변수일 때에는 LASSO를 로지스틱 회귀모형(Logistic regression model)에 적용할 수 있다. 2.2절에서 언급하였듯이, 로지스틱 회귀모형은 종속변수가 두 가지 값만 취하는 질적인 이분형 변수일 때, 확률에 대해 로짓 변환(logit transformation)을 한 것이다.

다시 살펴보면, p개의 독립변수에 대해서 종속변수가 **1**을 가질 확률을 $P(Y=1|x_1, x_2, ..., x_p)$라고 할 때, 로지스틱 반응함수는 다음과 같이 정리할 수 있다.

$$P(Y=1|x_1, x_2, ..., x_p) = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_p x_p)}$$

로지스틱 반응함수의 계수들을 추정하기 위해 최대우도추정(Maximum Likelihood Estimation, MLE)을 사용하는데, LASSO는 β에 대해 제약 조건이 있으므로 회귀계수 추정은 다음과 같이 나타낼 수 있다(Duffy and Santner,1989).

$$\arg\min l(\beta)$$

$$subjet\ to \sum_j |\beta_j| \leqq t$$

$l(\beta)$는 로그우도함수(log-likelihood function)이고, λ는 조절모수이다. 이 식은 다시 다음과 같이 나타낼 수 있다.

$$l^{\lambda}(\beta) = l(\beta) - \lambda|\beta|$$

$|\beta| = \sum_{j}|\beta_j|$ 이며, λ는 라그랑즈 승수(Lagrange multiplier)로, λ와 일대일 대응되는 조절모수이다. LASSO에서 회귀계수의 최대우도 추정은 뉴튼-랩슨 방법(Newton-Raphson method)이나 반복재가중최소제곱(iteratively reweighted least squares, IRLS)과 같은 반복적인 방법(iterative method)에 의해서 해의 근사값을 구하며, 3.4절의 알고리즘을 이용하여 제약조건의 수만큼 반복하는 과정을 거쳐 LASSO의 로지스틱 회귀모형의 해를 추정할 수 있다.

4장 건강검진 자료를 이용한 간경변 발생 판별 모형

4.1 건강검진 자료

4.1.1 소개

건강검진(Screening Test)은 종합적인 건강 상태를 파악하고 질병의 유무를 확인하는 검사로 질병의 조기진단 및 조기치료에 목적을 두고 있다. 본 논문은 1994년 5월 ~ 2005년 9월 사이에 이루어진 총 124,121건의 건강검진 자료를 바탕으로 하였다. 총 124,121건에는 최대 7회까지 건강검진을 받은 사람도 있으며, 이 가운데 가장 최근에 건강검진을 받은 85,458명을 대상으로 간경변 발생군에 대한 연구를 시작하였다.

4.1.2 건강검진 항목

건강검진센터에서 검진하고 있는 기본 검사 항목들은 기초정보, 신체계수, 폐기능 검사, 간 기능 검사, 안과 검사, 청력 검사, 심전도 검사, 내시경 검사, 영양상태, 혈액검사 등이 있으며, 문진은 2000년부터 건강검진 항목에 추가되었다. 다음은 건강검진센터에서 검진하는 세부항목을 나타낸 표이다.

표**1**. 건강검진 검사항목

검사항목	변수		검사항목	변수
기초정보	성별 SEX 연령 age		대사 및 전해질	나트륨 Na 칼륨 K 염소 Cl 이산화탄소 CO_2 칼슘 Ca 인 P 혈당 Glucose 당화혈색소 HbA_1C 혈중요소질소 BUN 크레아티닌 Creatinine 요산 Uric Acid
신체계측	신장 Height 체중 Weight BMI			
혈액검사	적혈구	RBC Hb Hct MCV MCH MCHC		
	백혈구	WBC 임파구 LYM 호산구 EOS 호염구 BAS	간염검사	B형간염S항원 HBsAg B형간염C항체 AntiHBc B형간염S항체 AntiHBs C형간염항체 AntiHCV B형간염E항원 HbeAg B형간염E항체 Anti-HBe
	혈소판	혈소판 Platelet		
	혈액형	ABO		
간 기능 검사	총단백 T.Protein 알부민 Albumin 총빌리루빈 T.Bilirubin 직접빌리루빈 D.bilirubin Alk. Phos AST ALT γ-GT LDH		뇨 검사	비중 SG 선도 pH 단백 Protein 요당 Glucose 케톤체 Ketone 잠혈 Blood 요빌리노겐 Urobilinogen 빌리루빈 Bilirubin 아질산염 Nitrite 백혈구 UWBC 색 Color 탁도 Turbidity
혈청지질	총콜레스테롤 T.cholesterol 중성지방 Triglyceride 고밀도콜레스테롤 HDL			
종양혈청	알파태아성단백 α-FP 태아성암항원 CEA		문진	흡연력 음주력 운동여부 가족력 정기적인 건강검진여부
내시경 검사	식도 Esophagus 위장관 Stomach Duodenum S상결장경검사 Sigmoid			

4.2 간경변 발생 판별 모형

4.2.1 분석 자료 소개

건강검진 센터에서 1994년 5월 ~ 2005년 9월 사이에 건강검진을 받은 85,458명 중 8,031명이 다시 병원에 내원하여 소화기내과 검진을 받았다. 적게는 1회에서 최대 12회까지 소화기내과 검진을 받았으며, 본 연구에서는 가장 최근의 검진 결과를 가지고 분석하였다.

간경변 발생 판별 모형을 위한 반응 변수는 '간경변 발생자'와 '간경변 비발생자'로 나누며, 8,031명 중 간경변 발생자는 978명이고 비발생자는 7,053명이다.

표2. 간경변 발생 분포

	빈도	퍼센트	총합
간경변 발생	978	12%	8,031
간경변 비발생	7,053	88%	

모형 추정을 위해 기존에 알려진 간경변 발생에 대한 위험 인자를 바탕으로 건강검진 항목에서 문진, 기초정보, 혈액검사, 간 기능 검사, 혈청 지질 검사, 대사 및 전해질 검사, 간염검사, 뇨 검사 항목이 독립변수로 사용되었다. 문진항목은 검사결과를 바탕으로 표3과 같이 음주력(Drinking), 운동여부(Exercise), 가족력(Family history), 정기적인 건강검진여부(Regular) 등 4개의 새로운 독립변수를 정의하였다. 음주력은 음주횟수, 술 종류, 회당 평균 음주량을 바탕으로 하루평균소주음주량을 계산하였고, 음주기간을 함께 고려하여 정의하였으며, 운동여부는 운동 횟수와 운동량을 고려하여 정의하였다. 가족력은 직계에 한해 간질환에 대한 가족력 여부를 나타내는 변수로 정의하였다.

표3. 간경변 발생에 대한 독립변수 정의

검사	독립변수	하위 검사 항목 결과
문진	Drinking	술 안 마신다.
		하루평균소주음주량 **1**병 미만이고 음주기간 **1**0년 미만.
		하루평균소주음주량 **1**병 이상이고 음주기간 **1**0년 미만이거나, 하루평균소주음주량 **1**병 미만이고 음주기간 **1**0년 이상.
		하루평균소주음주량 **1**병 이상이고 음주기간 **1**0년 이상.
	Exercise	운동 안 한다.
		주4회 미만이고 회당 60분 미만.
		주4회 이상이고 회당 60분 이상.
	Family history	간질환 가족력 없다.
		간질환 가족력 있다.
	Regular	정기적인 건강검진을 받지 않는다.
		매년 **1**회 건강검진을 받는다.
		매년 2회 이상 건강검진을 받는다.

문진을 통해 만들어진 4개의 변수를 포함하여 간경변 발생을 판별하기 위해 총 54개의 독립변수가 사용되었으며, 결측치를 제외한 4,093명이 분석에 사용되었다. 4,093명 중 간경변 발생자는 50**1**명이고 비발생자는 3,592명이다.

표4. 결측치를 제외한 간경변 발생 분포

	빈도	퍼센트	총합
간경변 발생	50**1**	**1**2%	4,093
간경변 비발생	3592	88%	

표5는 독립변수 54개에 대한 일변량 분석 결과를 나타낸 표이다.

표5. 독립변수의 일변량 분석

변수	비발생(n=3592) $mean \pm SD$	발생(n=501) $mean \pm SD$	$p-value$	
age	50.80±12.14	49.90±11.20	0.0947	
WBC	6.23±1.79	5.88±1.78	<.0001	**
RBC	4.46±0.52	4.48±0.51	0.4443	
Hb	13.98±1.63	14.32±1.49	<.0001	**
Hct	41.16±4.83	41.99±4.52	0.0003	**
MCV	92.34±5.26	93.90±5.54	<.0001	**
MCH	31.39±2.21	32.07±2.34	<.0001	**
MCHC	33.98±0.97	34.14±1.01	0.0010	**
LYM	37.55±8.44	38.88±9.63	0.0035	**
EOS	0.44±1.91	0.43±1.67	0.8247	
BAS	0.05±0.22	0.06±0.21	0.5772	
Platelet	243.64±62.39	199.09±71.01	<.0001	**
T.Protein	7.26±0.41	7.28±0.43	0.3177	
Albumin	4.52±0.30	4.38±0.38	<.0001	**
T.Bilirubin	0.81±0.41	0.96±0.78	<.0001	**
Alk. Phos	74.12±26.72	87.08±50.46	<.0001	**
AST	26.17±37.98	56.41±129.67	<.0001	**
ALT	29.15±60.67	62.57±143.22	<.0001	**
γ-GT	40.74±72.72	95.68±208.64	<.0001	**
LDH	348.97±128.33	375.80±115.36	<.0001	**
T.Cholesterol	192.95±35.07	184.13±36.99	<.0001	**
Triglycerides	148.75±102.65	133.01±118.58	0.0048	**
HDL	53.09±13.27	53.34±14.66	0.7300	
CEA	2.48±4.33	7.87±114.71	0.2940	
αFP	2.51±1.56	20.36±140.31	0.0046	**
Na	141.93±2.09	141.69±2.04	0.0185	*
K	4.22±0.35	4.17±0.35	0.0075	**
Cl	102.60±2.34	102.66±2.58	0.6336	
CO_2	26.19±2.39	26.19±2.30	0.9808	
Ca	9.64±0.45	9.52±0.46	<.0001	**
P	3.60±0.51	3.49±0.54	<.0001	**
Glucose	97.63±26.67	96.33±26.16	0.3056	
BUN	14.04±3.80	13.81±3.75	0.2132	
Creatinine	0.98±0.22	0.99±0.19	0.1353	
Uric Acid	5.09±1.39	5.06±1.29	0.5747	
SG	1.02±0.02	1.02±0.01	0.0237	*
pH	5.72±0.91	5.69±0.89	0.5345	
Urobilinogen	0.13±0.23	0.20±0.45	0.0008	**

* p-value < 0.05 , ** p-value < 0.01

표5. 독립변수의 일변량 분석 (계속)

변수		비발생(n=3592) 빈도(%)	발생(n=501) 빈도(%)	$p-value$	
Protein		381(10.61%)	66(13.17%)	0.0844	
Glucose		101(2.81%)	15(2.99%)	0.8179	
Ketone		168(4.68%)	40(7.98%)	0.0016	**
Blood		1154(32.13%)	143(28.54%)	0.1063	
Bilirubin		112(3.12%)	48(9.58%)	<.0001	**
Nitrite		25(0.70%)	4(0.80%)	0.7979	
UWBC		435(12.11%)	67(13.37%)	0.4195	
HBsAg		151(4.20%)	245(48.90%)	<.0001	**
AntiHBc		2117(58.94%)	403(80.44%)	<.0001	**
AntiHBs		2388(66.48%)	202(40.32%)	<.0001	**
AntiHCV		27(0.75%)	65(12.97%)	<.0001	**
SEX(Male)		1919(53.42%)	329(65.67%)	<.0001	**
Drinking	0	1580(43.99%)	151(30.14%)	<.0001	**
	1	214(5.96%)	16(3.19%)		
	2	1250(34.80%)	170(33.93%)		
	3	548(15.26%)	164(32.73%)		
Exercise	0	1968(54.79%)	296(59.08%)	0.0917	
	1	1082(30.12%)	137(27.35%)		
	2	542(15.09%)	68(13.57%)		
Family history		198(5.51%)	243(48.50%)	<.0001	**
Regular	0	1913(53.26%)	288(57.49%)	0.0864	
	1	1617(45.02%)	205(40.92%)		
	2	62(1.73%)	8(1.60%)		

* p-value < 0.05 , ** p-value < 0.01

일변량 분석 결과, WBC, Hb, Hct, Platelet, Albumin, T.Cholesterol, AST, ALT, Alk. Phos, αFP, HBsAg, AntiHBc, AntiHBs, AntiHCV, Drinking, Family history 등 33개의 변수가 간경변 발생군과 비발생군에서 유의한 차이가 있는 것으로 나타났다.

5장 간경변 발생 예측 모형

5.1 간경변 발생 예측 모형

간경변 발생 예측 모형을 통해 간경변 위험인자(risk factor)를 찾고, 이를 통해 예측력을 알아볼 수 있다. 본 논문에서는 4,093명을 대상으로 33개의 독립변수를 사용하였으며 변수들의 측정 기준이 다르므로 평균 0, 분산 **1**로 표준화(Standardization)한 후 예측 모형 연구를 시작한다. LASSO를 사용하며 LDA, QDA, 단계적 로지스틱 회귀분석, 의사결정나무, Random Forests, SVM-RFE와 그 성능을 비교해 볼 것이다.

모형의 성능을 평가하기 위해서 민감도와 특이도의 개념을 사용하기로 한다. 민감도(sensitivity)는 실제 '간경변 발생자'를 예측 모형에서도 간경변 발생자로 판단하는 비율이고, 특이도(specificity)는 실제 '간경변 비발생자'를 예측 모형에서도 비발생자로 판단하는 비율이다. False positive는 실제 '간경변 비발생자'를 예측 모형에서 간경변 발생자로 잘못 판단하는 비율로 '**1**-특이도'로 계산되며, false negative는 실제 '간경변 발생자'를 예측 모형에서 간경변 비발생자로 잘못 판단하는 비율로 '**1**-민감도'로 계산된다.

표6. 오분류표

		Actual	
		negative	positive
Predict	negative	specificity	false negative
	positive	false positive	sensitivity

예측 모형에서 민감도와 특이도가 모두 높을수록 좋은 모형이지만 실제로 두 비율이 동시에 높아지는 것은 불가능하다. 본 논문은 질병에 관한 연구이므로, 간경변 발생자를 예측모형에서도 간경변 발생자로 판단하는 비율인 민감도를 높이는 것이 중요하다. 한편, 모형의 성능을 평가하기 위해 4,093명의 자료를 7:3의 비율

로 훈련용 자료와 검증용 자료로 나누어 모형의 성능을 비교하고, 위험인자를 살펴보기로 한다.

5.2 변수 선택

6가지 판별방법에 따라 간경변 발생에 대한 위험인자로 선택된 변수들을 살펴보면 다음과 같다.

5.2.1 판별분석

다음의 표7은 판별 분석에 의해 선택된 변수를 정리한 표이다.

표7. 판별분석에서의 변수 선택

variable	F value	Pr > F	Wilks' Lambda	Pr > Lambda
HBsAg	881.37	<.0001	0.7646	<.0001
Family history	456.30	<.0001	0.6595	<.0001
AntiHCV	294.86	<.0001	0.5979	<.0001
$\gamma - GT$	75.13	<.0001	0.5826	<.0001
αFP	14.54	0.0001	0.5725	<.0001
Triglyceride	13.73	0.0002	0.5694	<.0001
Alk.Phos	10.78	0.0010	0.5676	<.0001
Platelet	11.59	0.0007	0.5653	<.0001
Albumin	7.53	0.0061	0.5638	<.0001
Hb	5.03	0.0250	0.5628	<.0001
Urobilinogen	4.20	0.0406	0.5620	<.0001
Drinking3	35.54	<.0001	0.5754	<.0001

판별분석에서 B형간염S항원(HBsAg), 가족력(Family history), C형간염항체(AntiHCV), $\gamma - GT$, αFP, 중성지방(Triglyceride), 알칼라인 포스파타제(Alk.Phos), 혈소판 수치(Platelet) 등 12개의 변수가 간경변 위험인자로 선택되었다.

5.2.2 단계적 로지스틱 회귀분석

표8. 단계적 로지스틱 회귀분석에서의 변수 선택

parameter	estimate	standard error	Wald Chi-Square	Pr>Chisq
HBsAg	3.3329	0.2318	206.69	<.0001
AntiHCV	3.7796	0.3283	132.55	<.0001
Family history	2.3865	0.1788	178.21	<.0001
Alk.Phos	0.2546	0.0616	17.0686	<.0001
Platelet	-0.2342	0.0799	8.59	0.0034
AntiHBs	0.4366	0.1970	4.9112	0.0267
Albumin	-0.1513	0.0728	4.3157	0.0378
Drinking 2	0.4116	0.1851	4.95	0.0262
Drinking 3	1.3819	0.1995	47.99	<.0001

단계적 로지스틱 회귀분석에서는 8개의 변수가 유의하게 나왔으며, B형간염S항원(HBsAg), C형간염항체(AntiHCV), 가족력(Family history), 알칼라인 포스파타제(Alk.Phos), 혈소판 수치(Platelet), B형간염S항체(AntiHBs), 알부민(Albumin), 음주력(Drinking)이 간경변 발생 위험인자로 나타났다. 판별분석과 마찬가지로 B형 및 C형 간염 바이러스, 가족력, 음주력이 유의하게 나타났다.

5.2.3 의사결정나무

CART 알고리즘을 사용하여 마디의 순수함을 나타내는 지니계수(Gini index)에 의해 분리기준(split criterion)을 정하였으며, 그림2는 CART에 의해 선택된 위험인자와 그것을 통해 형성된 간경변 판별 나무모형을 보여준다.

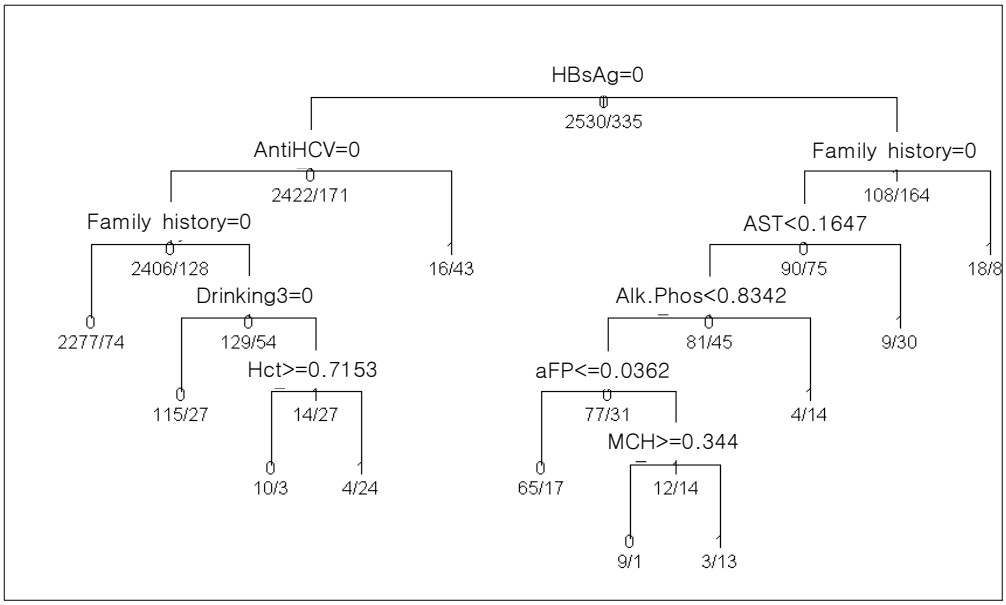

그림2. 의사결정나무에서의 변수 선택

B형간염S항원(HBsAg), C형간염항체(AntiHCV), 가족력(Family history), 음주력(Drinking), AST, 알칼라인 포스파타제(Alk.Phos), 알파태아성단백(αFP), Hct, MCH 등 9개의 변수가 선택되었다. 그림2를 보면, B형간염S항원(HBsAg)과 C형간염항체(AntiHCV)가 음성이고 가족력(Family history)이 없는 경우 간경변에 걸릴 위험이 적은 반면, B형간염S항원(HBsAg)이 양성이고, 가족력(Family history)이 있는 경우 간경변에 걸릴 위험이 큰 것으로 보인다. 다른 끝마디에 대해서도 이와 같은 해석을 할 수 있다.

5.2.4 Random Forests

Random Forests에서 변수의 중요도를 판단하기 위해 이 논문에서는 지니계수(Gini index)의 상대적인 감소율 정도로 중요 변수를 평가하였다.

표9. Random Forests에서의 변수 선택

variable	Mean DecreaseGini	variable	Mean DecreaseGini
HBsAg	71.2725	Hct	16.5539
Family history	51.5544	Hb	14.7139
αFP	41.1037	MCHC	14.1431
Platelet	36.6583	Ca	13.6009
AST	31.1216	P	12.7878
AntiHCV	26.9658	K	12.0336
$\gamma - GT$	21.3935	T.Bilirubin	11.8286
Triglyceride	20.8826	Na	10.3072
ALT	20.8495	Drinking	9.4961
Alk.Phos	20.6891	AntiHBs	6.6341
Lymph	17.7078	SG	6.0668
WBC	16.5492	AntiHBc	3.8366
Albumin	16.4870	SEX	1.4737
MCV	16.4383	Ketone	1.3311
MCH	16.3075	Bilirubin	1.1929
T.Cholesterol	16.2356	Uro-bilinogen	1.1879
LDH	16.2266		

Random Forests에서는 대부분의 변수들의 지니계수 감소율이 크게 나타났으며 그 중에서 B형간염S항원(HBsAg), 가족력(Family history), 알파태아성단백(αFP), 혈소판수치(Platelet), AST, C형간염항체(AntiHCV), $\gamma - GT$, 중성지방(Triglycerides), ALT, 알칼라인 포스파타제(Alk.Phos) 등 10개의 변수가 상대적으로 지니계수 감소율의 정도가 크게 나타났다.

5.2.5 SVM-RFE

SVM의 가중치를 이용하여 변수의 중요도를 구하는 SVM-RFE에서 모형에 사용될 변수의 수를 결정하기 위해 변수의 비율을 10%씩 변화해가면서 10차 교차타당성(10-fold cross-validation)을 하였고, 가장 작은 예측오차를 갖는 변수의 비율을 모형에 적용하였다.

표10. SVM-RFE에서의 변수 선택

변수	중요도(순위)
Family history	**1**
HBsAg	2
Drinking	3
AntiHCV	4
αFP	5
LDH	6
Platelet	7
WBC	8
$\gamma - GT$	9
Alk.Phos	**1**0

변수의 비율이 30%일 때 가장 작은 예측오차를 가지며, 표**1**0과 같이 중요도 순으로 **1**0개의 변수가 선택되었다. 가족력이 간경변에 가장 큰 영향을 미치며, 다른 판별방법과 마찬가지로 B형간염S항원(HBsAg), C형간염항체(AntiHCV)와 음주력(Drinking) 등이 간경변 발생에 대한 주요 위험인자로 나타났다.

5.2.6 LASSO

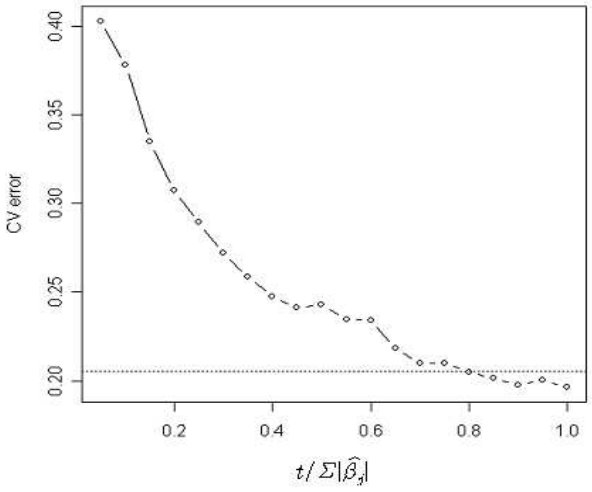

그림3. 조절모수 t의 결정

조절모수 λ를 결정하기 위해 10차 교차타당성(10-fold cross-validation)을 하였으며 추정된 오분류율의 $1-SE$ 원칙을 사용하여 그림3과 같이 λ를 결정하였고, 그때 LASSO에 의해 유의하게 선택된 변수는 표11과 같다.

표11. LASSO에서의 변수 선택

variable	estimate	variable	estimate
WBC	0	Na	0
Hb	0	K	0
Hct	0	Ca	0
MCV	0	P	0
MCH	0	SG	0
MCHC	0	Uro-bilirubin	0
Lymph	0	Ketone	0
Platelet	-0.1823	Bilirubin	0
Albumin	-0.0726	HBsAg	2.2931
T.Bilirubin	0	AntiHBc	0
Alk.Phos	0.1199	AntiHBs	0
AST	0	AntiHCV	2.5724
ALT	0	sex	0
$\gamma-GT$	0.0410	Family history	1.7156
LDH	0	Drinking 1	0
T.Cholesterol	0	2	0
Triglyceride	0	3	0.4011
αFP	0		

혈소판 수치(Platelet), 알부민(Albumin), 알칼라인 포스파타제(Alk.Phos), $\gamma-GT$, B형간염S항원(HBsAg), C형간염항체(AntiHCV), 가족력(Family history), 음주력(Drinking) 등 8개의 변수가 간경변 위험인자로 선택되었다. 회귀계수를 보면 혈소판 수치는 -0.1823, 알부민은 -0.0726으로 간경변이 발생하면 혈소판과 알부민이 감소한다는 것을 잘 나타내고 있다. 또한, B형간염항원(HBsAg)과 C형간염항체(AntiHCV)가 양성인 경우 그렇지 않은 경우에 비해 간경변이 발생할 위험이 더 크며, 가족력이 있는 경우에도 그렇지 않은 경우에 비해 간경변 발생 위험이 더 크다는 것을 알 수 있다. 음주력 또한 간경변에 영향을 주는 변수로서, 하루 평균

1병 이상의 소주를 10년 이상 마신 경우 그렇지 않은 경우에 비해 간경변 발생 위험이 더 크다. 8개의 변수를 제외한 나머지 변수들은 회귀계수가 0이므로 간경변 발생에 의미 없는 변수라고 할 수 있다.

다음은 훈련용 자료에서 각각의 판별방법을 통해 위험인자로 선택된 변수를 정리한 표이다.

표12. 간경변 발생에 대한 위험인자

model	Risk factor	
Discriminant Analysis	HBsAg, AntiHCV, Family history, Drinking, Alk.Phos, Hb, Platelet, Albumin, $\gamma-GT$, αFP, Triglyceride, Urobilinogen	(12개)
Logistic	HBsAg, AntiHCV, Family history, Drinking, Alk.Phos, Platelet, Albumin, AntiHBs	(8개)
TREE	HBsAg, AntiHCV, Family history, Drinking, Alk.Phos, αFP, AST, Hct, MCH	(9개)
Random Forests	HBsAg, AntiHCV, Family history, Alk.Phos, Platelet, $\gamma-GT$, αFP, Triglyceride, AST, ALT	(10개)
SVM-RFE	HBsAg, AntiHCV, Family history, Drinking, Alk.Phos, Platelet, $\gamma-GT$, αFP, LDH, WBC	(10개)
LASSO	HBsAg, AntiHCV, Family history, Drinking, Alk.Phos, Platelet, Albumin, $\gamma-GT$	(8개)

판별분석이 12개의 변수로 가장 많이 선택되었고, 단계적 로지스틱 회귀분석과 LASSO가 8개로 가장 적게 선택되었다. 선택된 변수를 살펴보면, 6개의 판별방법에서 B형간염S항원(HBsAg), C형간염항체(AntiHCV), 가족력(Family history), 알칼라인 포스파타제(Alk.Phos)가 공통으로 간경변 발생에 대한 위험인자로 선택되었다. Random Forests를 제외한 판별방법에서 음주력(Drinking)이 위험인자로 선택되어 음주여부도 간경변 발생에 영향을 주는 것으로 나타났으며, 의사결정나무를 제외한 판별방법에서는 혈소판 수치(Platelet)가 위험인자로 선택되었다. 이것은 간경변 발생 위험인자로 잘 알려진 B형 및 C형 간염 바이러스, 가족력, 음주력을 잘 나타내주며, 간 기능 검사를 통해서 간경변 발생을 예측할 수 있음을 보여준다.

변수 선택을 통해 모형을 비교한 결과, Random Forests나 SVM-RFE는 모형의 중요성만을 평가할 수 있는 반면, 판별분석이나 로지스틱 회귀분석, LASSO는 변수의 해석이나 유의성 검정이 가능하고, 모수절약의 원칙(parsimonious principle)을 고려할 때 모형이 훨씬 간결하다. 또한, LASSO는 의미 없는 변수의 회귀계수를 0으로 만들어주므로 변수의 해석이 쉽다.

5.3 간경변 발생 예측 모형 결과

다음의 표13과 표14는 각각 훈련용 자료와 검증용 자료에서 간경변 발생 예측 모형 결과를 정리한 표이다.

표13. 간경변 발생 예측모형 - 훈련용 자료

model	sensitivity	specificity	false negative	false positive	total accuracy	mis classification
LDA	68.38	94.87	31.62	5.13	91.62	8.38
QDA	56.98	96.46	43.02	3.54	91.62	8.38
Logistic	43.88	98.81	56.12	1.19	92.39	7.61
decision TREE	63.58	97.87	36.42	2.13	93.86	6.14
Random Forests	100	100	0	0	100	0
SVM-RFE	65.70	98.25	34.30	1.75	94.35	5.65
LASSO	60.7	96.62	39.30	3.38	92.25	7.75

표14. 간경변 발생 예측모형 - 검증용 자료

model	sensitivity	specificity	false negative	false positive	total accuracy	mis classification
LDA	58.67	95.55	41.33	4.45	91.04	8.96
QDA	52.67	96.47	47.33	3.53	91.12	8.88
Logistic	40.96	98.96	59.04	1.04	91.12	8.88
decision TREE	56.02	97.36	43.98	2.64	91.78	8.28
Random Forests	57.19	97.41	42.91	2.59	92.59	7.41
SVM-RFE	54.78	97.57	5.22	2.43	92.10	7.90
LASSO	58.27	96.14	41.73	3.86	91.60	8.40

검증용 자료의 모형을 비교해보면, 정확도는 Random Forests, SVM-RFE, 의사결정나무, LASSO 순으로 높게 나타났다. 민감도는 LDA, LASSO, Random Forests, 의사결정나무 순으로 높게 나타났으나 LDA, LASSO, Random Forests, 의사결정나무의 민감도가 비슷함을 알 수 있다. 또한, LASSO는 훈련용 자료와 검증용 자료에서의 민감도가 60.7%와 58.27%로 비슷한 성능을 보여 간경변 발생을 예측하는 데 안정적임을 보여준다.

5.4 LASSO를 이용한 간경변 발생 예측

LASSO의 훈련용 자료에서 위험인자로 나타난 B형간염S항원(HBsAg), C형간염항체(AntiHCV), 가족력(Family history), 음주력(Drinking3), 알칼라인 포스파타제(Alk.Phos), 혈소판 수치(Platelet), 알부민(Albumin), $\gamma - GT$ 등 8개의 변수를 다른 판별방법에 적용하여 간경변 발생을 예측하였으며 그 결과는 다음과 같다.

표15. LASSO에 의해 선택된 변수를 이용한 예측모형 - 훈련용 자료

model	sensitivity	specificity	false negative	false positive	total accuracy	mis classification
LDA	69.65	95.24	30.35	4.76	92.15	7.85
QDA	66.88	92.02	33.12	7.98	90.12	9.88
Logistic	45.38	98.73	54.62	1.27	92.29	7.71
decision TREE	60.98	97.74	39.02	2.26	93.30	6.70
Random Forests	69.36	99.72	30.64	0.28	96.06	3.94
SVM-RFE	64.98	98.68	35.02	1.32	94.49	5.51
LASSO	60.70	96.62	39.30	3.38	92.25	7.75

표16. LASSO에 의해 선택된 변수를 이용한 예측모형 - 검증용 자료

model	sensitivity	specificity	false negative	false positive	total accuracy	mis classification
LDA	57.42	94.69	42.58	5.31	89.98	10.02
QDA	50.02	91.53	49.98	8.47	87.82	2.18
Logistic	41.29	98.70	58.71	1.30	91.45	8.55
decision TREE	54.84	97.39	45.16	2.61	92.02	7.98
Random Forests	52.90	97.67	47.10	2.33	92.02	7.98
SVM-RFE	52.08	97.42	47.92	2.58	92.10	7.90
LASSO	58.27	96.14	41.73	3.86	91.60	8.40

LASSO에서 위험인자로 선택된 8개의 변수를 다른 판별방법에 적용한 결과, 특이도와 정확도 모두 높게 나타났다. 검증용 자료에서의 LDA의 민감도를 보면 57.42%로 LASSO의 민감도 58.27%와 비슷한 성능을 보인다. 그리고 의사결정나무와 Random Forests, SVM-RFE도 각각 54.84%, 52.90%, 52.09%의 민감도를 보여, LASSO에서 유의하게 선택된 8개의 변수만으로도 기존에 8개 이상의 변수를 사용한 분석결과와 비슷한 결과를 보임을 알 수 있다.

6장 결론 및 고찰

지금까지 건강검진 자료를 통한 간경변 발생을 예측하기 위해 LASSO를 사용하였으며, 그 성능을 기존의 판별방법인 LDA, QDA, 단계적 로지스틱 회귀분석, 의사결정나무, Random Forests, SVM-RFE와 비교하였다.

본 연구에서는 건강검진을 받은 검진자 중 병원에 내원하여 소화기내과 검사를 받은 4,093명을 대상으로 간경변 발생 예측 모형을 연구하였다. 즉, LASSO를 통해 검진자료만을 가지고 간경변 발생을 얼마나 잘 예측할 수 있는지, 또 간경변 발생 위험인자가 무엇인지 파악하고자 하였다. 그 결과, LASSO에 의한 간경변 발생 예측 모형은 91.6%의 정확도와 58.27%의 민감도로 다른 판별기법과 비슷한 성능을 보였으며, 간경변 발생에 대한 위험인자는 B형간염S항원(HBsAg), C형간염항체(AntiHCV), 가족력(Family history), 음주력(Drinking), 알칼라인 포스파타제(Alk.Phos), 혈소판 수치(Platelet), 알부민(Albumin), $\gamma-GT$ 등 8개의 변수로 나타났다. 이것은 기존에 알려진 간경변 발생 위험인자(risk factor)인 B형 및 C형 간염 바이러스, 가족력, 음주력을 포함하여 건강검진 항목의 간 기능 검사를 통해 간경변 발생을 58.27% 예측할 수 있음을 의미한다. 다른 판별방법들을 통해서도 B형 및 C형 간염과 가족력, 음주력 뿐만 아니라 간 기능 검사, 혈청 지질 검사 등을 통해 간경변을 예측할 수 있는 것으로 나타났다. 또한, LASSO의 결과에서 유의한 8개의 변수를 제외한 나머지 변수들의 회귀계수는 0이 되는데, 이것은 LASSO의 제약조건 $\sum_j |\beta_j| \le t$으로 인해 의미 없는 변수들의 회귀계수를 완전히 0으로 만들어줌으로써 예측력(prediction accuracy)을 높여주고, 쉽게 해석할 수 있는(interpretable) 모형을 제공해주는 LASSO의 특성을 잘 보여준다. 따라서 LASSO는 판별과 동시에 변수선택이 필요한 자료에 탁월하다는 것을 알 수 있다.

본 연구를 토대로 건강검진자료를 통해 간경변에 대한 조기진단을 마련할 수 있을 것으로 보이며, 추후에는 반복적인 검진자료를 통해서 간경변을 예측해 볼 수 있을 것이다. 또한, LASSO는 판별방법뿐만 아니라 비례위험모형(proportional

hazards model)과 같은 다른 여러 모형에도 적용 가능하므로 향후에는 LASSO의 생존분석(survival analysis)에 관한 연구도 이루어지길 바란다.

참 고 문 헌

김부성, 간경변증, 대한소화기학회 총서 6, 대한소화기학회, 군자출판사

한요셉, 김병호, 간암의 조기 진단을 위한 검진대상, 대한소화기학회지, 2001

한은정, 건강검진 자료에서 Random Forests를 이용한 백내장 발생 위험군 예측 모형, 2004

김소연, 혼합 로지스틱 분포를 이용한 백내장 발생 예측 모형 연구, 2005

오은주, 리찌와 라쏘 회귀분석법의 비교와 신용 점수화에의 적용, 2005

성웅현, 응용 다변량분석, 탐진, 2002

김동건, 조권익, SVM을 이용한 변수선택법 구현. 정보과학연구, 2003:7;107-120

David W. Hosmer, Stanley Lemeshow. Applied Logistic Regression. Wiley, 2002

Hastie.T., Tibshirani.R., Friedman.J. The Elements of Statistical Learning. Springer, 2001

Quinlan, J R. Introduction to decision tree. *Machine Learning*, 1986:1;81-106

Breiman.L, J H Freidman, R., A Olshen, and C J Stone. Classification and Regression Tree. *Belmont Wordsworth*, 1984;56-82

Breiman.L, Random Forests. *Machine Learning*. 2001:45(1);5-32

Tibshirani.R. Regression and Shrinkage via lasso. *Journal of the Royal Statistical Society.(Series B).* 1996:58(1);267-288

Wenjiang J. Fu. Penalized Regressions:The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics*. 1998:7(3);397-416

M.Osborne, B.Presnell, B.Turlach. A new approach to variable selection in least squares problems. *IMA J. Numerical Analysis*, 2000:20;389-404

Debashis Ghosh, Arul M.Chinnaiyan. Classification and selection of biomarkers in genomi data using LASSO. *Journal of Biomedicine and Biotechnology*, 2005:2;147-154

Knut Baumann. Chance correlation in variable subset regression : Influence of the objective function, the selection mechanism, and ensemble averaging. *QSAR & Combinatorial Science*, 2005:9;1033-1046

ABSTRACT

A study of prediction model for the development of liver cirrhosis using LASSO

Kim, Sung Eun

Dept. of Biostatistics and Computing

The Graduate School

Yonsei University

Liver cirrhosis can induce a complication, so it is necessary to reduce the incidence of liver cirrhosis by early diagnosis such as screening test.

This study investigated prediction model for the development of liver cirrhosis and identified the risk factors related with liver cirrhosis based on screening data accumulated from **1**994 to 2005. We applied the LASSO to the logistic regression model(LASSO-logistic regression), used it as prediction model and compared the performance of any other methods with ours. As a result, accuracy and sensitivity using LASSO-logistic regression were 9**1**.6%, 58.27%. These indicated that LASSO-logistic regression is comparable to performance of any others. And, we found that the risk factors related with liver cirrhosis were HBsAg, AntiHCV, Family history, Drinking, Platelet, Alkaline phosphatase, Albumin and $\gamma - GT$. This shows that these are well known risk factors of liver cirrhosis. Also, when using LASSO-logistic regression, it is easy to interpretate the results in contrast to decision tree, Random Forests, SVM-RFE. Therefore, LASSO-logistic regression is suitable for data analysis to have need

of variable selection and discriminant analysis simultaneously.

Key words : Liver cirrhosis, Screening test, Risk factor, Prediction model, Sensitivity, Accuracy, shrinkage, variable selection, LASSO