간경변증 발생 위험 예측 모형 구축을 위한 PLTR의 적용

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
임 용 진

간경변증 발생 위험 예측 모형 구축을 위한 PLTR의 적용

지도 김 도 영 교수

이 논문을 석사 학위논문으로 제출함

2008년 12월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

임 용 진

임용진의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2008년 12월 일

감사의 글

대학원에 입학한지 2년이라는 시간이 지나고 어느덧 졸업을 하게 되었습니다. 대학원 생활을 하면서 논문을 완성하기까지 많은 도움과 힘이 되어 주신 분들께 감사의 마음을 전하고자 합니다.

우선 바쁘신 와중에도 저의 논문 심사를 맡아주신 김도영 선생님께 감사드립니다. 통계라는 학문을 처음 접하면서 부족한 저에게 4년간의 대학생활 동안 통계학이 무엇인지 깨닫게 해주신 박동권 교수님, 하은호 교수님, 문명상 교수님, 나성룡 교수님께 감사드립니다. 많은 관심을 보여주신 남정모 교수님, 대학원에 입학을 하고 졸업하기까지 진정한 대학원 생활이 무엇인지 알려주시고 의학통계라는 학문을 공부하는데 있어서 많은 도움을 주신 송기준 선생님, 항상 같은 연구실에서 사적으로나 공적으로 많은 도움과 힘을 주신 명성민 선생님께 진심으로 감사드립니다. 2년동안 통계학의 깊이를 더해주신 김동건 교수님, 조진남 교수님, 임길섭 박사님께도 감사드립니다.

대학원 생활을 하면서 변함없이 따뜻하게 감싸주신 무형이형, 원열이형, 은희누나, 졸업 후에도 많은 도움을 주신 정윤누나, 많은 조언을 아끼지 않으셨던 임현선 선생님에게 고마운 마음을 전합니다. 항상 옆에서 챙겨주면서 힘이 되어주고 정말 재미있는 대학원 생활을 하게 해준 고마운 수희누나, 영애누나, 경화누나, 진희누나, 친 누나 같은 성유누나, 정말 착한 동기 혜선누나, 많은 시간을 같이 했던 낙훈이형, 부족한 저를 잘 따라준 성훈씨, 성희씨, 수연이, 대학원 후배지만 항상 밝은 나의 영원한 든든한 선배 성혁이형이 있어서 대학원 생활이 즐거웠고 잘 마무리한 것 같습니다.

단짝과도 같은 나의 든든한 버팀목 요한이형, 항상 지켜봐주면서 응원해주던 너무도 고마운 정민이, 희경이누나, 철영이, 민수, 수광이, 대학생활부터 지금까지 함께 지내며 많은 웃음을 준 사랑스런 친구 혜성이, 민영이, 나라, 함께 있으면 항상 유쾌하고 믿음직스러운 재현이형, 효경이형, 재영이형, 지수형, 창훈이형, 만재형에

게 곁에 있어줘서 늘 행복하다고 전하고 싶습니다. 나의 소중한 평생지기 덕근이, 해선이, 문모, 한나, 금주, 인선누나에게 늘 고마운 마음뿐입니다.

멀리 떨어져있지만 항상 기도로서 응원해주고 지켜주신 아버지, 어머니, 한나에게 사랑한다는 말을 전하고 싶습니다. 사랑하는 든든한 가족이 있어 제가 여기까지 올 수 있었습니다. 마지막으로, 보이지 않는 곳에서 늘 곁에서 지켜주시는 하나님 아버지께 감사드립니다.

2008년 12월

임용진 올림

차 례

표 차 례

그림 차례

국 문 요 약

간경변증 발생 위험 예측 모형 구축을 위한 PLTR의 적용

질병에 위험을 주는 다양한 특성에 대해 알아보기 위해 많은 통계학적 방법들이 이용되어지고 있는데, 특히 회귀분석방법과 나무모형을 이용한 방법들이 주로 이용되고 있다. 하지만 이런 방법들은 다중 변수 조합의 효과를 평가하는데 있어 고려하는 요인의 수가 증가하게 되면, 그에 따른 가능한 상호작용의 수도 기하급수적으로 증가하기 때문에 많은 요인을 다루는 능력에 있어서 한계를 보이게 된다. 이에 대한 대안으로 일반화선형모형과 나무모형의 이점들을 결합시킨 회귀모형에 기초한 부분적 선형 나무 (Partially Linear Tree-based Regression model, 이하 PLTR) 모형이 제안(Jinbo Chen et al., 2006)되었다. 이 방법은 선형모형과 비모수적 나무모형의 조합을 이용하여 위험요인들의 주 효과와 결합 효과를 동시에 효과적으로 나타내줄 수 있다.

본 논문에서는 PLTR을 이용하여 간경변증 발생 위험 예측 모형에 대한 유용성을 평가하고자 건강검진 센터에서 1994년 5월부터 2005년 9월 사이에 검진을 받은 검진자 중 다시 병원에 내원하여 간경변증 발생 여부에 대한 진단을 받은 4,093명의 자료를 이용하였다. 평가를 위해 로지스틱 회귀분석, QUEST, SPAN등의 방법을 적용하여 민감도, 특이도, 정확도를 통하여 비교 하였다.

PLTR에 의해 선택된 간경변증 발생 위험인자는 HBsAg, Anti-HCV, family history, history of drinking, platelet, alkaline phosphate로 기존에 알려진 간경변 발생 위험인자를 대체적으로 잘 나타내주고 있었다. 로지스틱 회귀분석과 QUEST에 비해 민감도와 정확도가 우월한 것으로 판단해 볼 때, PLTR은 예측 모형을 구축하는데 그 유용성이 존재함을 확인할 수 있었다.

핵심되는 말 : PLTR, 로지스틱회귀분석, QUEST, SPAN, 민감도, 특이도, 정확도

제 1장 서론

1.1 연구배경

간경변증이란 만성적인 염증으로 인해 정상적인 간 조직이 바뀌어 간의 기능이 저하되는 것을 의미한다. 간경변증의 원인은 다양한데, 만성 B형 간염이나 C형 간염, 지속적인 과음과 간 독성 물질의 사용 등으로 간의 염증상태가 지속되는 경우에 간경변증이 발생하게 된다. 이러한 간경변증의 발생을 예측하고 위험군 분류를 위한 통계학적 방법들이 이용되고 있다. 대표적으로 회귀분석 방법 중에서 로지스틱 회귀분석을 통하여 B형 및 C형 간염 바이러스, 음주력, 가족력 관련 요인들이 간경변증 발생 위험인자로 나타나는 것을 확인 할 수가 있었다. 간경변증 발생 위험 예측 모형 구축을 위해 유전 분야에서 발달된 방법을 적용한다.

질병에 위험을 주는 다양한 특성에 대해 알아보기 위해 유전 분야에서도 통계학적 방법들이 이용되어지고 있다. 유전자와 같은 다중 변수(multiple variable) 조합의 효과를 평가하는데 있어 고려하는 요인의 수가 증가하게 되면, 그에 따른 가능한 상호작용의 수도 기하급수적으로 증가하기 때문에 많은 요인을 다루는 능력에 있어서 한계를 보이게 된다(하정윤, 2007). 일반적으로 회귀모형이 주 효과와 저차 유전자-유전자 상호작용(gene-gene interaction)의 효과를 평가하는데 있어서 편리하게 이용되어지고 있지만, 복잡한 고차 상호작용의 효과를 평가하기에는 어려움이 있다. 또한 데이터마이닝의 한 분야로 나무모형을 이용한 분석 방법 또한 빈번하게 이용되어지고 있다. 이러한 나무모형은 복잡한 고차 상호작용을 다루는데 있어서는 좋은 방법이 되지만, 주 효과를 평가하는데 있어서는 어려움이 있다. 이러한 단점들을 보완하고자 일반화선형모형과 나무모형의 이점들만 나타내주는 방법인 회귀모형에 기초한 부분적 선형 나무 (Partially Linear Tree-based Regression model, 이하 PLTR) 모형이 제안(Jinbo Chen et al., 2006)되었다. 이 방법은 선형모형과 비모수적 나무모형의 조합에 의해 위험요인들의 주 효과와 결합 효과를 동시에 효과적으로 나타내준다.

1.2 연구목적 및 내용

본 논문에서는 PLTR 모형을 통하여 간경변증 발생 위험 요인분석에도 적용이 가능한지 알아보고자 한다. PLTR 모형의 적용은 환자에 대한 일반적 특성과 검사항목에서 간경변증 발생에 선형적으로 유의한 영향을 주는 요인들을 선형모형으로 나타내고 나머지 요인들은 나무모형으로 나타내서 그들의 조합으로 PLTR 모형에 적용시킨다.

PLTR 모형과의 비교를 위해 고전적인 통계 방법인 로지스틱 회귀분석, 데이터 마이닝의 방법으로서 나무모형에서 일반적으로 쓰이고 있는 Classification and Regression Tree(이하 CART, Breiman et al., 1984)의 단점을 보완한 Quick Unbiased Efficient Statistical Tree(이하 QUEST, Loh and Shih, 1997), 나무모형의 제한점들을 보완하여 제안되어진 Search Partition Analysis(이하 SPAN, Roger Marshall, 1986)의 방법을 이용하였다. 로지스틱 회귀분석은 SAS v9.1.2(SAS Institute, Inc), QUEST는 QUEST v1.9.2, SPAN은 SPAN package, 그리고 본 논문에서 제안하는 방법인 PLTR 모형은 R-package의 rpart library와 SAS v9.1.2(SAS Institute, Inc)를 사용하여 각 방법들에 대한 결과를 도출하였다. 위 방법들에 대하여 민감도, 특이도, 정확도의 평가 방법들을 통하여 임상적 위험요인 분석에서 PLTR 모형의 효과적인 효율성을 판단하고자 한다.

논문의 구성은 1장 서론에서 연구의 배경과 목적 및 내용에 대해서 언급하였고, 제 2장에서는 비교를 위한 연구 방법들인 로지스틱 회귀분석, QUEST, SPAN을 설명하고 제 3장에서는 PLTR 모형의 이론적 내용을 설명하였다. 제 4장에서는 건강검진 센터에서 1994년 5월부터 2005년 9월 사이에 검진을 받은 검진자 중 다시 병원에 내원하여 간경변증 발생 여부에 대한 진단을 받은 4,093명으로 구성되어있는 자료에 대한 설명과 각 방법들에 대한 비교 결과들에 대하여 설명하였다. 마지막으로 제 5장에서는 본 논문에 대한 결론 및 고찰에 대해서 논의하였다.

제 2장
이분형 자료의 분류 방법에 대한 통계적 고찰

2.1 로지스틱 회귀분석

2.1.1 이론적 배경

종속변수가 두 가지 값만을 취하는 이분형 질적 변수일 때, 선형회귀모형에서 설정하는 가정을 심하게 위반하게 된다. “성공”, “실패” 등과, 두 개의 값을 가지는 이분형 질적 종속변수와 독립변수와의 관계를 설명하기 위한 회귀분석의 방법 중 대표적인 통계분석 방법이 로지스틱 회귀분석 인데, 로지스틱 회귀모형은 선형 회귀모형의 제한적인 가정들을 극복할 수 있다.

2.1.2 로지스틱 회귀 모형

로지스틱 회귀 모형은 아래와 같이 변수를 정의 한다.

Y_i : 종속변수, 어떤 사건이 발생 한 경우를 1, 발생하지 않은 경우를 0

x_i : 각 관찰치 값(observation) , $i=0,1,2,3,\cdots,n$

β_i : 회귀계수(regression parameter) , $i=0,1,2,3,\cdots,k$

$P(Y_i=1|x_i)$: 관찰치 x_i에 대해 종속변수가 1을 가질 확률이라고 정의

모형의 형태가 선형을 만족하도록 로짓변환(logit transformation)을 시키면

$$\log \frac{P(Y_i=1|x_i)}{1-P(Y_i=1|x_i)} = \beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ni}$$

이다. 여기서 모수 β_i에 대해서 일차 선형관계를 갖고 있음을 알 수 있다. 위의 식을 $P(Y_i=1|x_i)$에 관하여 정리하면

$$P(Y_i=1|x_i)=\frac{\exp(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ni})}{1+\exp(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ni})}$$

이 되고, 이 식을 로지스틱 반응함수(logistic response function)이라 한다.

2.1.3 로지스틱 회귀계수의 추정과 검정

로지스틱 회귀모형에서 모수인 회귀계수를 추정하는 데 가장 널리 사용되는 방법은 최대우도추정법(maximum likelihood estimation)이다. 이 때 우도(likelihood)는 관찰된 자료가 발생될 확률을 알려지지 않은 모수들의 함수로 표현한 것이며, 이 우도를 최대화 시키는 추정법이 최대우도추정법이다. 이렇게 얻어진 모수들의 추정량을 최대우도추정량(maximum likelihood estimator)이라 한다. 로그우도함수는 아래와 같다.

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i \ln(p) + (1-y_i)\ln(1-p)]$$

여기서 $p=\frac{1}{1+\exp[-(\beta_0+\beta_1 X)]}$ 인데, 이를 β_0와 β_1에 대해 편미분하여 정규방정식(normal equation)을 구하면 다음과 같다.

$$\sum_{i=1}^{n}(y_i-\hat{p})=0, \quad \sum_{i=1}^{n}x(y_i-\hat{p})=0$$

이 방정식을 Newton-Raphson 알고리즘을 이용하여 최대우도추정량을 구할 수 있다.

이렇게 우도함수의 최대값은 미분을 통해 얻게 되는데, 회귀계수의 최대우도 추정값은 비선형이므로 피셔의 스코어링방법(Fisher's method of scoring)이나 뉴튼-랩슨 방법(Newton-Raphson method)등과 같은 반복적인 추정방법에 의하여 근사값를 구한다(성웅현, 2001).

회귀계수를 추정한 후에 로지스틱 회귀모형에 대한 검정은 우도비 검정(likelihood ratio test), 왈드 검정(Wald test), 스코어 검정(score test) 등을 이용한다.

2.2 QUEST

2.2.1 이론적 배경

나무 모형 중에서 일반적으로 사용되는 CART(Classification and Regression Tree, Breiman et al., 1984)는 자식마디를 형성할 때 보다 많은 이산값을 가지는 예측변수를 선택하는 경향이 있기 때문에, 계산시간이 다소 많이 걸리고 분류 또는 예측오차가 커질 가능성이 있다. 그러므로 변수선택 편의(bias)나 계산시간을 줄이고자 하는 방법으로 QUEST(Quick Unbiased Efficient Statistical Tree, Loh and Shih, 1977)가 제안되었다.

2.2.2 변수 선택

QUEST를 수행하게 위해서는 변수선택을 먼저 시행하는데, 변수선택 알고리즘을 요약하면 다음과 같다(Loh and Shih, 1977).

1. 순서형 예측변수에 대해서는, 분산분석을 수행하여 F검정의 p값을 계산한다.
2. 범주형 예측변수에 대해서는, 예측변수와 목표변수의 분할표에서 카이제곱검정의 p값을 계산한다.
3. 1,2 단계에서 가장 작은 p값이 Bonferroni의 수정된 임계값(critical value)보다 작으면 그에 대응되는 변수를 분리변수로 선택한다.
4. 3단계에 해당 되지 않는 경우, 순서형 예측변수에 대하여 Levene F검정의 p값을 계산하고 Bonferroni의 수정된 임계값과 비교한다. p값에 대응되는 변수를 분리변수로 선택한다.

2.2.3 분리 규칙

QUEST는 명목형 목표변수에 대해서만 분석을 수행할 수 있으며, 예측변수의 측도에 따라서 서로 다른 분리규칙을 사용한다. 예측변수가 순서형 또는 연속형인 경우에는 분리규칙(splitting rule)으로 분산분석 F검정 또는 Levene의 검정을 사용하며, 명목형인 경우 Pearson의 카이제곱 검정을 사용한다. 목표변수의 범주가 3개 이상인 경우에는 k-평균 군집분석(k-means clustering)에서 군집의 수 k가 2개인 2-평균 군집분석(two-means clustering)을 수행하여 두 개의 그룹을 만든 후 분석을 수행한다(Loh and Shih, 1977). 또한 예측변수의 최적분리를 찾기 위해 2차 판별분석(quadratic discriminant analysis)을 수행하고, 목표변수를 가장 잘 분류하는 예측 변수의 최적의 분리를 이용하여 자식마디(Child node)를 형성하게 된다. 이와 같이 QUEST에서는 목표변수에 대한 예측변수의 기여도를 F검정 또는 카이제곱검정의 p값으로 측정함으로서, 변수선택의 편의를 줄이도록 하였다. 따라서 관측치의 수가 많거나 복잡한 자료에 대해서는 효율적이라고 할 수 있다.

2.2.4 모형 선택

위의 두 단계를 반복적으로 실시하면서 나무 모형을 만들고 각 단계마다 오분류 비용(missclassification cost)를 계산한다. 오분류 비용이 정해놓은 값보다 작은 나무 모형을 최종모형으로 선택한다.

2.3 SPAN

2.3.1 이론적 배경

최근 다양한 요인을 분류하는 방법으로 나무모형이 빈번이 이용되어 지고 있다. 나무모형을 이용한 방법은 자료를 세분화하는 과정에서 동질적인 하위그룹을 얻는 방법이다. 하지만 나무모형을 이용한 방법은 위계적인 것으로 결과를 적용하는데 제한점이 있다(Marshall RJ, 2001). 이에 대한 대안으로 SPAN(Search Partition Analysis)이 제안 되었다.

2.3.2 속성

SPAN은 자료를 두 개의 그룹으로 나누는 알고리즘이다. 둘로 나뉜 분류의 결과는 속성들의 부울 결합(Boolean combination)을 이용하여 표현된다. 여기서 속성이랑 질병에 대한 위험인자로 생각할 수 있는 것들이다. 속성들의 부울 결합은 자료를 두 개의 공간으로 나누게 되는데, 이때의 목표는 두 개의 공간을 가장 동질적으로 만드는 분류결과를 찾는 것이다.

속성 정의를 한 후에 $\mathcal{S}$를 잘 반영하는 m개의 속성을 선택하여 속성들의 집합인

$$T_m=\{X_1, X_2, \cdots, X_m\}$$

을 정한다. 이때의 m의 크기는 연구자가 임의로 선택한다. 보통 m은 카이제곱 통계량을 이용하여 통계적으로 유의한 속성의 수로 결정한다.

2.3.3 부울 결합

T_m의 크기인 m을 정하고 나면, 부울 결합의 크기를 정한다. 부울 결합들 중에 S, S' 를 각각을 가장 동질적으로 하는 부울 결합을 찾는 것은 불순도의 기준인

$$I(P) = -P\log P-(1-P)\log(1-P)$$

를 사용한

$$G= I(P_S)-P_A I(P_{S|A})-P_{A'} I(P_{S|A'})$$

를 이용한다. 여기서 P_A와 $P_{A'}$는 A와 A'의 확률이고, $P_{S|A}$와 $P_{S|A'}$는 A와 A'가 주어졌을 때 S의 조건부 확률이다. 정해진 m q p_i에 대한 모든 가능한 경우의 수만큼의 부울 결합에 대해 G값을 모두 구해 이를 가장 크게 하는 부울 결합을 찾는다.

2.3.4 모형 크기의 제한

최상의 부울 결합을 찾는 과정에서 그 크기가 무한히 커지는 것을 막기 위해

$$G-c\beta$$

를 이용하여 제한을 주는데, 이때의

$$c=q+q'-1$$

이고 β는 c가 하나씩 늘어남에 따라 나타나는 기울기로 이때 기울기의 증가 하는 정도가 작아지면 그 전단계의 c에서 G를 최대로 하는 부울 결합을 선택한다.

제 3장 회귀모형에 기초한 부분적 선형 나무모형

3.1 이론적 배경

유전학 분야의 방대한 자료에서 복합성 질환(complex diseases)에 영향을 주는 다양한 요인들의 특성을 확인하기 위해 회귀모형에 기초한 부분적 선형 나무 모형 (Partially Linear Tree-based Regression, 이하 PLTR) 모형이 제안되었다. 이 방법은 일반화선형모형과 나무모형의 이점들만 나타내 주는 모형으로 linear-term과 tree-term을 모두 포함하여 주 효과와 상호작용을 동시에 효과적으로 고려해 준다.

3.2 속성

PLTR 모형은 주 효과와 상호작용의 효과를 동시에 고려해주는 모형이다. 선형 모형 부분의 주 효과로서 적용 시키고 싶은 변수들의 집합을 X로 하고, 나무모형 부분에서 상호작용의 효과로서 적용 시키고 싶은 변수들의 집합을 $G=(G_1, G_2, \ldots\ldots, G_M)$로 한다.

이분형 Y에 대해 $E(Y)=P(Y=1)=g(\eta)$라 가정하자. 이때 g는 로지스틱 연결함수(link function) $g(\eta)=\exp(\eta)/1+\exp(\eta)$이다. 이러한 반응변수 Y에 X와 G의 결합 효과를 표현하기 위해 제안된 PLTR 모형은 다음과 같이 표현된다.

$$\eta=\gamma' X+\beta' Z(T)$$

위험 요인 X는 선형으로 표현되고 G의 상호작용 효과는 나무구조 T를 이용

하여 표현한다. 이때, $A(T)$는 나무구조 $T(G)$에서의 마지막 노드를 나타내주는 행렬이다. 관찰치 j가 j번째 노드로 떨어질 때의 값을 Z^{jj}라 하고 그 외의 경우에는 0으로 놓게 된다.

3.3 반복 적합 알고리즘

Clark and Pregibon(1993)이 편차(deviance)에 기초한 반복적인 배분 방법을 제안했다. 이 방법은 n_0, n_1이 각각 정상과 대조군의 수를 나타낼 때 노드의 편차를 $n_0\log(n_0)+n_1\log(n_1)$ 로 정의한다. 후보자 노드에 대한 최적의 배분은 부모 노드의 편차에서 두 자식 노드의 편차들의 합을 뺌으로서 알 수 있다. 여기서 자식 노드의 두 후보자로 나눌 때, 0/1의 가변수 Z로 표현되는 반응변수 Y에 대한 로지스틱 회귀에서의 편차 통계량이라는 것을 쉽게 볼 수 있다.

$\gamma'X$를 로지스틱 회귀모형에 포함되는 예측변수로 간주하고 PLTR 모형에 적합시키기 위해 이용한다. j번째 개체에 대해 동일한 예측변수가 $\gamma'X$가 나무의 모든 노드에 적용된다.

이러한 내용을 바탕으로 편차 함수를 최소화시켜주는 반복 적합 알고리즘(iterative fitting algorithm)을 제안한다(Jinbo Chen et al., 2006).

step 1. 기본적인 나무모형 T와 상응하는 모형으로, Z가 1인 single column으로 표현되는 $\eta=\gamma'X+\beta'A(T)$에 적합 시킨다.

step 2. 다음 ①, ②, ③ 단계의 반복 작업

①. $\widehat{\gamma'X}$를 로지스틱 회귀모형에 포함되는 예측변수로 간주하고 나무모형 $T(G)$에 적합 시킨다.

②. *step 1*①.에서의 나무모형을 기반으로 $\eta = \hat{\gamma}'X + \beta'Z(T)$에 다시 적합 시킨다. 앞에서와 같이 $\hat{\gamma}'X$는 예측변수로 이용되고, 추정치 β를 $\hat{\beta}$로 다시 표시한다.
③. $\hat{\beta}'Z(T)$를 예측변수로 이용하고 $\eta = \gamma'X + \hat{\beta}'Z(T)$적합 시킴으로서 γ를 업데이트 시킨다.

step 3. 알고리즘은 γ의 추정치가 특정한 범위 안에서 안정되어질 때 멈추게 된다.

이렇게 적합 된 나무모형은 데이터를 과 적합하여 지나치게 커질 수가 있다. 이러한 경우 적절한 크기로 가지치기(pruning)를 하게 된다.

3.4 최적 나무의 선택과 유의성 평가

최적 나무의 선택과 이에 대한 유의성을 평가하기 위해 Yu et al.(2005)에 의해 제안된 단계-전진 탐색 알고리즘(step-forward search algorithm)을 이용한다. 이 알고리즘은 Draper and Smith(1981)의 단계-전진 변수 선택법(step-forward variable selection)과 유사한 방법이다. 앞서 설명한 반복 알고리즘에 의해 적합된 모형을

$$\ln[\gamma = 1|X, G] = \hat{\gamma}'X + \beta'Z[T_j(G)] \qquad (2)$$

라 하자. 이때 $T_j(G)$는 수렴하는 나무모형으로 표현되고 j는 나무모형 마지막 노드의 수를 뜻한다. $T_j(G)$가 과적합(over fitting) 되었을 때 G의 결합 효과를 잡아주기 위한 최적의 나무모형을 찾는 것을 목적으로 한다.

이에 대한 기본적인 개념은 i번째 마지막 노드에서 특정 정수를 나타내는 m에서의 $T_j(G)$에 지분된 후보자의 하위나무의 배열을 정의하는 것이다.

$T_i(G), i=1, \ldots\ldots, m$

$\ln[\gamma=1|X, G]=\hat{\gamma}^0 X$ 모형에 상응하는 편차를 $D(X, \hat{\gamma}^0)$ 로 정의하고 (2) 모형에 상응하는 편차는 $T_f(G)$가 하위나무 $T_i(G)$로 표현 때문에 $D(X, T_i(G), \hat{\gamma}^1]$로 정의한다.

첫 번째, 이 알고리즘은 나무 T_f 안에서 $\Delta D_i = D(X, \hat{\gamma}^0) - D(X, T_2^i(G), \hat{\gamma}^1]$의 편차가 가장 크게 감소하는 하위나무들을 포함하면서 T_1에서 하나로 나뉘는 것에서부터 시작된다.

두 번째, 각 하위나무 $T_i, i=1,2,\cdots, m$에서의 각각의 경험적인 유의확률을 계산해준다. 이때 $p^* = \min(p_i, i=1,2,\cdots, m)$가 되는 최적의 나무 T_i^*를 선택한다.

세 번째, 최적의 나무모형이 반응 변수 Y와 유의한 관련이 있는지 평가를 한다.

제 4장 간경변증 발생 위험 예측 모형 구축을 위한 PLTR의 적용

4.1 자료에 대한 개요

건강검진 센터에서 1994년 5월부터 2005년 9월 사이에 이루어진 총 124,121건의 건강검진 자료를 바탕으로 중복된 검진자들은 가장 최근의 정보를 담아 총 85,458명을 추출하였다. 이들 중 다시 병원에 내원하여 소화기내과 검진을 받은 8,031명의 자료를 구성하였다. 2000년부터 실시한 문진항목을 추가하고, 결측치가 존재하는 대상은 제외하여 최종적으로 4,093명을 분석 자료로 사용하였다. 4,093명 중 간경변 발생분포를 살펴보면 간경변 발생자는 501명이고 비발생자는 3,592명으로 나타났다. 본 논문에서는 모형을 구축하고, 그 모형을 검증하기 위하여 4,093명 자료를 임의로 훈련용 자료와 검증용 자료로 나누어 분석을 실시하였다. 훈련용 자료에서의 간경변 발생자는 250명이고, 검증용 자료에서의 간경변 발생자는 251명으로 [표 1]과 같다.

표 1. 간경변 발생 분포

	전체자료	훈련용 자료	검증용 자료
간경변 비발생	3,592(87.76%)	1795(87.73%)	1797(87.79%)
간경변 발생	501(12.24%)	251(12.27%)	250(12.21%)
계	4,093(100%)	2046(100%)	2047(100%)

4.2 분석에 쓰인 건강검진 항목

분석을 위해 기존에 알려진 간경변 발생 위험 인자를 바탕으로 건강검진 항에서 기초정보, 혈액 검사, 간기능 검사, 혈청지질 검사, 종양혈청 검사, 대사 및 전해질 검사, 간염 검사, 뇨 검사, 문진 항목을 사용하였다. 다음 [표 3]는 검진항목의 세부내용을 정리한 것이다.

표 2. 건강검진 세부항목(계속)

검사항목	변수	영문	한글
기초정보	sex		성별
	age		연령
혈액검사	RBC	red blood cell	적혈구
	Hb	hemoglobin	헤모글로빈
	Hct	hematocrit	헤마토크리트
	MCV	mean corpuscular volume	평균적혈구용적
	MCH	mean corpuscular hemoglobin	평균적혈구혈색소량
	MCHC	mean corpuscular hemoglobin concentration	평균적혈구혈색소농도
	WBC	white blood cell	백혈구
	LYM	lymphocyte	림프구
	EOS	eosinocyte	호산구
	BAS	basophil leukocyte	호염구
	platelet	platelet	혈소판
대사 및 전해질	Na	sodium	나트륨
	K	potassium	칼륨
	Cl	chlorine	염소
	CO_2	carbon dioxide	이산화탄소
	Ca	calcium	칼슘
	P	phosphorus	인
	glucose	blood glucose	혈당
	BUN	blood urea nitrogen	혈중요소질소
	creatinine	creatinine	크레아티닌
	uric acid	uric acid	요산

표 2. 건강검진 세부항목

검사항목	변수	영문	한글
간기능검사	T.protein	total protein	총단백
	albumin	albumin	알부민
	T.bilirubin	total bilirubin	총빌리루빈
	Alk.phos	alkaline phophatase	알칼리성 포스파타제
	AST	aspartate aminotransferase	아스파르테이트 아미노전이효소
	ALT	alanine aminotransferase	알라닌아미노전이효소
	r-GT	gamma-glutamyl transferase	감마-글루타밀전이효소
	LDH	lactic dehydrogenase	젖산탈수소효소
혈청지질	T.cholesterol	total cholesterol	총콜레스테롤
	triglyceride	triglyceride	중성지방
	HDL	high-density lipoprotein	고밀도콜레스테롤
간염검사	HBsAg	hepatitis b virus	B형간염 바이러스
	Anti-HBc		B형간염 C항체
	Anti-HCV	antihepatitis C virus	C형간염 항체
종양혈청	α-FP	alpha-fetoprotein	알파-태아단백
	CEA	carcinoembryonic antigen	태아성암항원
뇨검사	SG	specific gravity	비중
	pH	hydrogen ion concentration	수소이온농도지수
	protein	protein	단백질
	urine glucose	urine glucose	요당
	ketone	ketone body	케톤체
	blood	occult blood	잠혈
	urobilinogen	urobilinogen	우로빌리노겐
	bilirubin	bilirubin	빌리루빈
	nitrite	nitrite	아질산염
	UWBC	white blood cell	백혈구
문진	family history		가족력
	drinking		음주력
	excercise		운동여부

4.3 자료의 일변량 분석

훈련용 자료(2,046명)를 이용하여 총 51개의 변수 중 38개의 연속형 변수의 일변량 분석결과는 [표 3]에 정리하였고, 13개의 이산형 변수에 대한 일변량 분석결과는 [표 4]에 정리하였다. 연속형 독립변수의 일변량 분석 결과 22개의 변수 age, Platelet, Hb, Hct, MCV, MCH, Wbc, LYM, Urobilinogen, Albumin, T.Bilirubin, Alk.phos, AST, ALT, γ-GT, LDH, Na, Ca, P, T.Cholesterol, triglycerides, α-FP, 이산형 독립변수의 일변량 분석 결과 6개의 변수 Bilirubin, HBsAg, Anti-HCV, sex, drinking, family history 가 유의한 차이를 보였다. 총 28개의 변수가 간경변 발생군과 비발생군 간에 유의한 차이가 있는 것으로 나타났다.

표 3. 연속형 독립변수의 일변량 분석 (훈련용 자료)

변수	발생(N=251) 평균 ±표준편차	비발생(N=1,795) 평균 ±표준편차	유의확률
age	49.29 ±11.49	50.99 ±11.97	0.0341*
Platelet	201.75 ±74.54	244.01 ±64.57	<.0001*
RBC	4.48 ±0.52	4.47 ±0.51	0.7144
Hb	14.31 ±1.57	14.01 ±1.60	0.0049*
Hct	42.00 ±4.75	41.22 ±4.75	0.0156*
MCV	93.94 ±6.12	92.40 ±5.35	0.0002*
MCH	32.06 ±2.59	31.42 ±2.25	0.0003*
MCHC	34.10 ±1.09	33.99 ±0.98	0.1189
Wbc	5.83 ±1.82	6.24 ±1.73	0.0005*
LYM	39.26 ±9.50	37.41 ±8.25	0.0036*
EOS	0.46 ±1.81	0.42 ±1.87	0.7511
BAS	0.07 ±0.24	0.05 ±0.24	0.3404
SG	1.02 ±0.01	1.02 ±0.03	0.1001
pH	5.66 ±0.88	5.68 ±0.87	0.7579
Urobilinogen	0.23 ±0.58	0.14 ±0.22	0.0104*
T.Protein	7.28 ±0.43	7.27 ±0.40	0.7943
Albumin	4.38 ±0.40	4.52 ±0.30	<.0001*
T.Bilirubin	0.95 ±0.65	0.81 ±0.40	0.0008*
Alk.phos	87.30 ±46.54	73.73 ±26.24	<.0001*
AST	53.81 ±85.89	25.27 ±17.70	<.0001*
ALT	56.75 ±74.90	27.78 ±29.67	<.0001*
γ-GT	104.78 ±227.90	41.80 ±82.48	<.0001*
LDH	376.51 ±124.03	349.23 ±158.24	0.0018*
Na	141.67 ±2.03	141.95 ±2.13	0.0443*
K	4.19 ±0.33	4.22 ±0.36	0.1618
Cl	102.67 ±2.51	102.59 ±2.37	0.6163
Co2	26.13 ±2.27	26.18 ±2.37	0.7794
Ca	9.51 ±0.48	9.63 ±0.45	<.0001*
P	3.49 ±0.50	3.59 ±0.54	0.0087*
glucose	97.88 ±28.23	97.32 ±25.52	0.7670
BUN	13.57 ±3.31	13.96 ±3.61	0.0846
creatinine	0.98 ±0.19	0.98 ±0.20	0.8700
Uric Acid	4.96 ±1.28	5.09 ±1.41	0.1648
T.Cholesterol	185.58 ±34.90	193.56 ±35.27	0.0008*
triglycerides	130.25 ±101.47	150.55 ±107.69	0.0049*
HDL	53.42 ±14.56	53.08 ±13.48	0.7075
CEA	2.76 ±2.93	2.61 ±4.89	0.5055
α-FP	29.83 ±191.89	2.49 ±1.45	0.0249*

표 4. 이산형 독립변수의 일변량 분석 (훈련용 자료)

변수	발생(N=251)	비발생(N=1,795)	유의확률
sex	158	956	0.0039
family history	120	99	<.0001*
drinking	76	264	<.0001*
exercise	150	971	0.0910
Protein	35	179	0.0541
Blood	69	588	0.0941
UWBC	36	226	0.4365
Nitrite	1	18	0.4990
Glucose	9	45	0.3180
Ketone	17	87	0.1932
Bilirubin	29	53	<.0001*
HBsAg	130	88	<.0001*
Anti-HCV	28	19	<.0001*

4.4 변수선택

간경변 발생 위험군을 분류하기 위해 간경변 발생의 위험인자를 찾고, 이를 통해 분류의 정확도를 파악하였다. 본 논문에서는 4,093명에서 22개의 연속형 독립변수와 6개의 이산형 변수들을 가지고 PLTR과 로지스틱 회귀분석, QUEST, SPAN의 성능을 비교하였다. 로지스틱 회귀 분석은 SAS v9.1.2(SAS Institute, Inc), QUEST는 QUEST v1.9.2, SPAN은 SPAN package, 그리고 본 논문에서 제안하는 방법인 PLTR 모형은 R-package의 rpart library와 SAS v9.1.2(SAS Institute, Inc)을 사용하여 각 방법들에 대한 결과를 도출하였다.

분류 결과에 대한 평가를 위해서 민감도와 특이도, 위양성율과 위음성율, 정확도(total accuracy)를 사용하였다. 민감도(sensitivity)는 실제 간경변 발생자 중 예측분류에서도 간경변 발생자로 분류되는 비율을 나타내는 것이고, 특이도(specificity)는 실제 간경변 비발생자 중 예측분류에서도 간경변 비발생자로 분류되는 비율을 나타낸다. 위양성율(false positive)은 실제 간경변 비발생자 중 간경변 발생자로 잘못 분류된 비율이고, 위음성율(false negative)은 실제 간경변 발생자 중 간경변 비발생자로 잘못 분류된 비율이다. 정확도(accuracy)는 실제 간경변 발생/비발생과 예측분류에서의 간경변 발생/비발생의 분류가 일치하는 정도를 의미한다. 각 방법들의 분류 능력을 평가하기 위해서 훈련용 자료와 검증용 자료로 나누어 살펴보았다.

4.4.1 로지스틱 회귀분석

단계적 로지스틱 회귀분석에서는 Alk.phos, AST, triglyceride, HBsAg, Anti-HCV, family history, drinking이 간경변 발생의 위험인자로 나타났다. 선택된 변수들의 계수와 유의확률은 [표 5]과 같다.

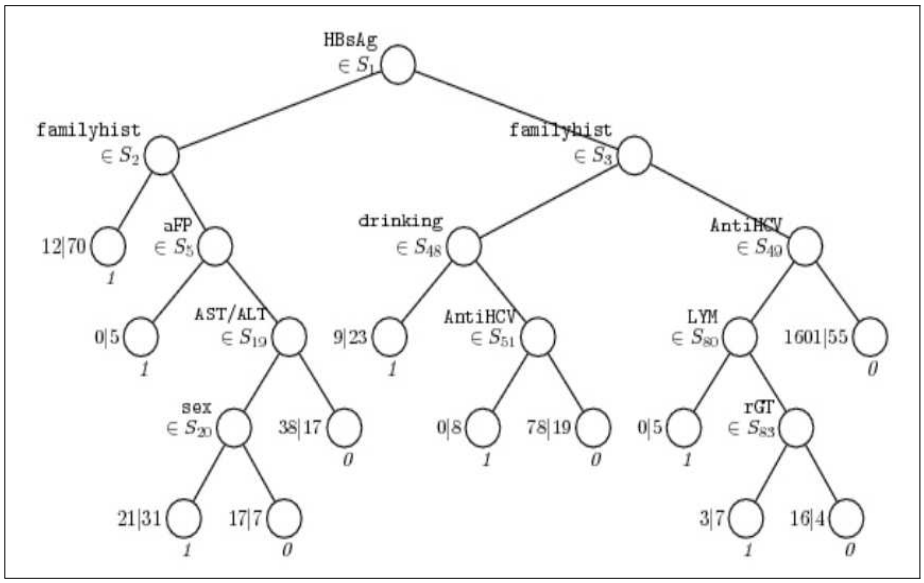
표 5. 단계적 로지스틱 회귀분석에서의 결과 (훈련용 자료)

변수	회귀계수	표준오차	유의확률
Alk.phos	0.006	0.002	0.0106
AST	0.011	0.003	0.0001
triglyceride	−0.003	0.001	0.0030
HBsAg	2.862	0.204	<.0001
Anti-HCV	3.219	0.365	<.0001
family history	2.394	0.209	<.0001
drinking	1.027	0.214	<.0001

4.4.2 QUEST

QUEST에서 사전확률로서 자료의 발생확률을 이용하고, 오분류비용은 같게 하고, 변수선택은 통계적 방법으로, 분리 기준은 지니지수(Gini criterion)로 하여 모형을 설정하였다. 간경변증 발생 위험인자는 HBsAg, family history, AST/ALT, sex, α-FP, drinking, Anti-HCV, LYM, γ-GT으로 9개가 선택되었다. [그림 1]은 QUEST에서 선택된 변수들의 나무모형으로 제일 왼쪽의 가지를 보면 HBsAg가 양성이고, family history가 있으면 간경변 위험군으로 분류되는 것을 보여준다.

그림 1. QUEST에서 선택된 변수들의 나무모형 (훈련용 자료)



4.4.3 SPAN

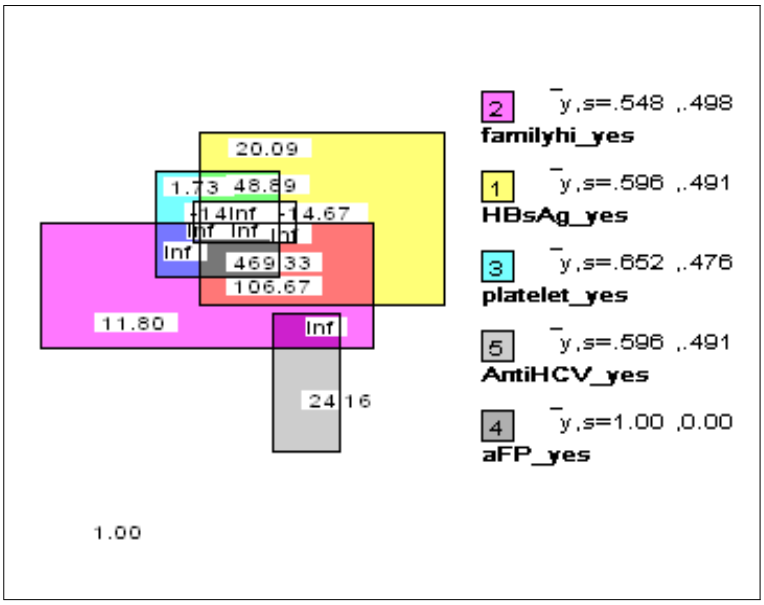
SPAN에서는 민감도와 특이도가 최대가 되는 지점인 변수 5개로 이루어진 모형을 선택하였다. 선택된 간경변증 발생 위험인자는 platelet, HBsAg, Anti-HCV, α-FP, family history로 나타났다. 이를 부울 결합을 이용하여 표현하면

간경변증 위험군 : B형간염 바이러스=양성 or 가족력=유 or 혈소판<130(10^3/uL) or 알파태아성단백>20(IU/L) or C형간염항체=양성

간경변증 비위험군 : B형간염 바이러스=음성 and 가족력=무 and 혈소판>=130(10^3/uL)and 알파태아성단백<=20(IU/L) and C형간염항체=음성

이다. [그림 2]은 SPAN에서 상자그림과 해당변수의 오즈비(Odds ratio)를 나타낸 것이다.

그림 2. SPAN에서 상자그림과 해당 오즈비 (훈련용 자료)



4.4.4 회귀모형에 기초한 부분적 선형 나무모형

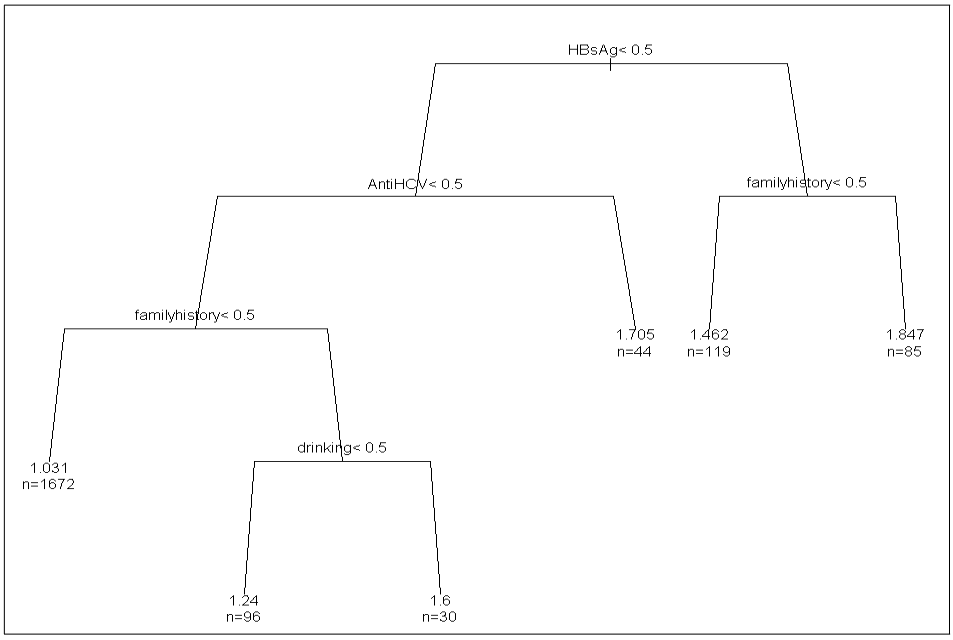
PLTR은 선형모형과 비모수적 나무모형의 조합에 의해 위험요인들의 주 효과와 결합 효과를 동시에 효과적으로 나타내주는 모형이다. 우선 PLTR 모형에서 선형 모형으로서 주 효과를 보기 위해 간경변 발생군과 비발생군에 선형적으로 유의한 영향을 미치는 독립변수들을 찾았다. 연속형 독립변수의 일변량 분석에서 유의한 차이를 보였던 22개의 변수들을 기초로 다중로지스틱 회귀분석을 통해 간경변 발생군과 비발생군에 선형적으로 유의한 영향을 미치는 독립변수들을 [표 6]에 정리하였다. 다중로지스틱 회귀분석분석을 통해 Hb, Platelet, Albumin, Alk.Phos, AST, triglycerides, 6개의 독립변수가 선택이 되었다.

표 6. 연속형 변수에 대한 다중로지스틱 회귀분석 (훈련용 자료)

변수	회귀계수	표준오차	유의확률
Hb	0.155	0.054	0.0042
Platelet	−0.012	0.001	<.0001
Albumin	−1.037	0.248	<.0001
Alk.Phos	0.005	0.002	0.0241
AST	0.004	0.001	0.0086
triglycerides	−0.003	0.001	0.0022

PLTR 모형에서의 나무모형으로서 결합 효과를 보기 위해 간경변 발생군과 비발생군에 유의한 영향을 미치는 이분형으로 되어있는 이산형 독립변수들을 찾았다. 이러한 변수들을 이용하여 PLTR 모형을 위한 나무모형 $\mathcal{T}(G)$를 설정하였다.

그림 3. PLTR 모형을 위한 나무 모형 $\mathcal{T}(G)$ (훈련용 자료)



PLTR 모형을 위한 나무모형 $\mathcal{T}(G)$에서 HBsAg, AntiHCV, family history, drinking 이 선택되었고 6개의 마지막 노드는 다음과 같다.

노드 1 : HBsAg=음성 & Anti-HCV=음성 & family history=무

노드 2 : HBsAg=음성 & Anti-HCV=양성 & family history=유 & drinking=무

노드 3 : HBsAg=음성 & Anti-HCV=양성 & family history=유 & drinking=유

노드 4 : HBsAg=음성 & Anti-HCV=양성

노드 5 : HBsAg=양성 & family history=무

노드 6 : HBsAg=양성 & family history=유

나무모형으로 설정되어진 $\pi(G)$를 보면 B형간염 바이러스로부터 시작이 되어 C형간염 항체와 가족력으로 분류가 되는 것을 볼 수 있다. PLTR 모형에서 나무모형으로서의 결합 효과를 보기 위해 B형간염 바이러스와 C형간염 항체가 음성이고 가족력이 없다고 분류되어진 노드 1을 기준으로 나머지 마지막 노드들을 가변수로 만든다.

PLTR 모형에서 linear-term으로서 주 효과를 보기 위해 간경변 발생군과 비발생군에 선형적으로 영향을 미치는 6개의 변수들과 tree-term으로서 결합 효과를 보기 위한 나무 모형 $\pi(G)$에서의 마지막 노드들과의 조합으로 이루어진 PLTR 모형은 [표 7]과 같다. 간경변증 발생에 대해 Platelet와 Alk.Phos의 주 효과를 볼 수 있고, B형 간염바이러스, C형간염 항체, 가족력, 음주력의 상호작용 효과를 볼 수가 있다.

표 7. PLTR 모형 (훈련용 자료)

변수	회귀계수	표준오차	유의확률
Platelet	−0.008	0.002	<.0001
Alk.Phos	0.009	0.002	<.0001
node2	2.460	0.287	<.0001
node3	3.900	0.421	<.0001
node4	4.333	0.375	<.0001
node5	3.058	0.244	<.0001
node6	4.792	0.346	<.0001

[표 8]은 앞에서 언급된 4가지 방법들에서 간경변 발생 위험 인자로 선택된 변수들을 정리한 것이다. 선택된 변수들을 살펴보면 HBsAg, Anti-HCV, family history가 모든 방법에서 선택되었고 SPAN을 제외한 3가지 방법에서는 drinking이 선택되었다. 그밖에 Alk.phos, platelet, triglyceride, AST, LYM, γ-GT, α-FP, sex가 선택되었다.

표 8. 모형별 변수선택을 통한 위험 인자

모형	위험 인자			
로지스틱 회귀분석	HBsAg	Anti-HCV	family history	drinking
	Alk.phos	AST	triglyceride	
QUEST	HBsAg	Anti-HCV	family history	drinking
	LYM	γ-GT	α-FP	sex
SPAN	HBsAg	Anti-HCV	family history	α-FP
	platelet			
PLTR	HBsAg	Anti-HCV	family history	drinking
	Platelet	Alk.Phos		

4.5 간경변증 발생 위험군 분류 결과

훈련용 자료를 통하여 4가지 모형에서 선택된 변수들을 가지고 간경변 발생에 대한 민감도(sensitivity), 특이도(specificity), 위음성율(False negative), 위양성율(False positive), 정확도(total accuracy)를 다음의 [표 9]에 정리하였다. 로지스틱 회귀분석에서 전체적인 특이도가 높았고 민감도는 낮았다. PLTR 모형이 비록 특이도는 로지스틱 회귀분석 보다는 낮지만 정확도가 높고 위음성율을 낮추기 때문에 PLTR 모형의 적절성을 확인 할 수 있다. 실제 간경변 발생자를 예측분류에서도 간경변 발생자로 분류하는 비율인 민감도는 SPAN에서 가장 높은 것으로 나타났고 PLTR 모형이 그 다음으로 높은 것으로 나타났다. [표 10]은 검증용 자료에서의 분석 결과로 훈련용 자료에서의 결과와 큰 차이를 보이지는 않았다. 여기에서도 PLTR 모형은 로지스틱 회귀분석과, QUEST 보다는 높은 민감도를 보였으나 SPAN보다는 낮은 민감도를 보였다. 정확도에서는 PLTR 모형이 높은 정확도를 보였다.

표 9. 간경변증 위험군 분류 결과 (훈련용 자료)

모형	민감도	특이도	위음성율	위양성율	정확도
로지스틱 회귀분석	48.2	98.4	51.8	1.6	92.3
QUEST	49.0	97.7	51.0	2.3	91.7
SPAN	78.5	88.2	21.5	11.8	87.0
PLTR	67.6	97.2	32.4	2.8	92.4

표 10. 간경변증 위험군 분류 결과 (검증용 자료)

모형	민감도	특이도	위음성율	위양성율	정확도
로지스틱 회귀분석	45.6	98.4	54.4	1.6	91.9
QUEST	51.2	97.9	48.8	2.1	92.2
SPAN	81.6	89.9	18.4	10.1	88.8
PLTR	64.2	97.9	35.8	2.1	93.7

4.6 공통변수 분석결과

PLTR 모형에서 선택된 HBsAg, Anti-HCV, family history, drinking, platelet, Alk.phos를 가지고 로지스틱 회귀분석, QUEST, SPAN, PLTR 방법으로 분석한 결과를 [표 11]과 [표 12]에 정리하였다. 훈련용 자료와 검증용 자료에서 PLTR 모형이 정확도가 가장 높았고, SPAN 보다 민감도가 낮았지만 로지스틱 회귀분석과 QUEST 보다는 민감도가 높았다.

표 11. PLTR에서 선택된 변수를 대상으로 한 분석결과 (훈련용 자료)

모형	민감도	특이도	위음성율	위양성율	정확도
로지스틱 회귀분석	50.2	97.5	49.8	2.5	91.7
QUEST	43.1	98.8	56.9	1.2	91.9
SPAN	82.5	90.1	17.5	9.9	89.2
PLTR	67.6	97.2	32.4	2.8	92.4

표 12. PLTR에서 선택된 변수를 대상으로 한 분석결과 (검증용 자료)

모형	민감도	특이도	위음성율	위양성율	정확도
로지스틱 회귀분석	49.8	97.9	50.2	2.1	92.0
QUEST	42.8	98.6	57.2	1.4	91.8
SPAN	85.4	91.6	14.6	8.4	90.8
PLTR	64.2	97.9	35.8	2.1	93.7

다음은 로지스틱 회귀분석, QUEST, SPAN, PLTR 모형에서 선택된 HBsAg, Anti-HCV, family history, drinking, Alk.phos, AST triglyceride, LYM, γ-GT, α-FP, sex, platelet 모든 변수들을 가지고 분석하였다. [표 13]을 보면 선택된 모든 변수를 대상으로 분석한 결과로 훈련용 자료에서 PLTR 모형이 정확도 92.4로 가장 높게 나타났으나 SPAN과 비교했을 때 민감도는 낮고 특이도는 높았다. [표 14]의 검증용 자료에서도 비슷한 양상으로 나타났다.

표 13. 선택된 모든 변수를 대상으로 한 분석결과 (훈련용 자료)

모형	민감도	특이도	위음성율	위양성율	정확도
로지스틱 회귀분석	43.4	98.3	56.6	1.7	91.6
QUEST	52.2	97.7	47.8	2.3	92.1
SPAN	78.4	88.1	21.6	11.9	86.9
PLTR	59.4	97.1	40.6	2.9	92.4

표 14. 선택된 모든 변수를 대상으로 한 분석결과 (검증용 자료)

모형	민감도	특이도	위음성율	위양성율	정확도
로지스틱 회귀분석	41.6	98.8	58.4	1.2	91.8
QUEST	48.3	97.1	51.7	2.9	91.1
SPAN	80.4	88.6	19.6	11.4	87.6
PLTR	63.0	98.2	37.0	1.8	93.9

5. 결론 및 고찰

유전 분야에서 다양한 특성들에 대한 유전자-유전자와 유전자-환경의 복잡한 조합을 확인함에 있어서 통계학적 방법들이 이용되어지고 있다. 이렇게 복잡한 조합의 효과를 평가하는데 있어서 고려하는 요인의 수가 증가하게 되면 그에 따른 가능한 상호작용의 수도 기하급수적으로 증가하기 때문에 많은 요인을 다루는 능력에 있어서 한계를 보이게 된다. 이러한 문제를 해결하고자 본 논문에서는 회귀모형과 나무모형에서의 이점들만 나타내주는 방법인 PLTR 모형을 제안하였다. PLTR 모형은 고려하는 요인의 수가 많을 때 요인에 대한 주 효과와 상호작용의 효과를 동시에 효과적으로 나타내주는 모형이다. 유전 분야에서 제안된 PLTR 모형이 임상적 위험요인 분석에도 효과가 있을 것으로 보여 로지스틱 회귀분석, QUEST, SPAN을 이용하여 간경변 발생 위험군 분류 결과를 비교해 보았다.

본 논문에서는 1994년 5월부터 2005년 9월 사이에 검진을 받은 검진자 중 다시 병원에 내원하여 간경변증 발생 여부에 대한 진단을 받은 4,093명의 자료를 이용하였다.

분석 결과 PLTR 모형은 나머지 3가지 방법들 중 로지스틱 회귀분석과 QUEST보다 실제 간경변 비발생자를 예측분류에서도 간경변 비발생자로 분류하는 특이도가 좀 더 낮은 결과를 보였고, 실제 간경변 발생자를 예측분류에서도 간경변 발생자로 분류하는 민감도에 있어서는 높은 결과를 보였다. SPAN보다는 낮은 민감도를 보였으나 특이도와 정확도는 높게 나타났다.

PLTR 모형에서 가장 주안점으로 보게 되는 부분이 선형모형과 나무모형의 결합으로 주 효과와 상호작용의 효과를 동시에 볼 수 있다는 점이다. 질병 발생에 선형적으로 강한 영향을 주는 요인들을 주 효과로 고정시킨 후 나머지 요인들과의 상호작용을 고려하여 주 효과와 상호작용 효과의 조합으로 최종적인 모형이 얻어 지게 되는 것이다. 임상적 위험요인 분석에서의 유용성 평가를 위해 로지스틱 회귀분석, QUEST, SPAN의 방법들과 비교를 했을 경우 월등히 높은 결과나 낮은 결과를 보이지는 않았다. 로지스틱 회귀분석에서는 복잡한 요인에 대한 상호

작용의 효과는 나타내고 있지 않고, QUEST나 SPAN의 경우에는 관측치의 수가 많거나 복잡한 자료에 대해서는 효율적인 방법이라고 할 수는 있지만 각 요인에 대한 주 효과는 나타내지 않는다는 제한점들이 있게 된다. 임상 분야에서의 효과적인 유용성을 비교해 보았을 때 PLTR 모형은 주 효과와 상호작용의 효과를 동시에 보여주면서 민감도, 특이도, 정확도에서 큰 차이를 보이지 않는다는 점으로 보아 유용할 것이라고 보인다.

질병 발생에 대하여 예측 또는 분류를 목적으로 하는 경우 실제 질병 발생자를 예측분류에서도 질병 발생자로 분류하는 민감도에 관심을 가진다. 분석 결과를 보면 로지스틱 회귀분석, QUEST, PLTR의 경우 각 방법에서 특이도보다 민감도가 떨어지는 결과를 보인다. 이는 간경변증 자료의 분포에서 간경변증 발생군이 비발생군 보다 현저히 적게 분포되어 있어서 이러한 결과를 보이는 것으로 생각해 볼 수 있다. 각 분석 방법들을 비교해 보았을 때 SPAN의 경우 민감도가 아주 높게 나타나는 것을 확인할 수가 있었다. SPAN은 두 개의 그룹으로 분류 하는 목적으로 하는 알고리즘인데 질병에 대한 위험인자들의 부울 결합(Boolean combination)을 이용하여 표현하게 된다. 애초에 최적의 분류를 목적으로 하기 때문에 높은 민감도를 보이게 된다. 하지만 PLTR은 SPAN과 비교해 보았을 경우 예측(prediction)이 가능한 모형으로 나타나기 때문에 적용이 더 이로울 것이다. 예측모형인 로지스틱 회귀분석과 비교해 보았을 경우 PLTR 모형은 로지스틱 회귀분석 보다 특이도는 낮지만 정확도와 민감도가 높고 위음성율을 낮추기 때문에 PLTR 모형의 적절성을 확인 하였다.

PLTR 모형을 임상 분야에 적용함에 있어서 주 효과와 상호작용 효과로 나타낼 요인들에 대한 명확한 구분이 필요 할 것으로 보인다. 이러한 알고리즘이 제안된다면 좀 더 효과적으로 임상 분야에서도 질병 발생을 예측 할 것으로 보여 진다.

참고 문헌

강현철. SAS Enterprise Miner 4.0을 이용한 데이터마이닝. 2001.

배화수. SAS Enterprise Miner를 이용한 데이터마이닝. 2005.

성웅현. 응용 로지스틱 회귀분석. 2001.

유영애. 간경변증 발생 위험군 분류를 위한 SPAN의 유용성 평가. 연세대 대학원 석사 학위논문, 2007.

하정윤. MDR을 이용한 간경변증 발생 고위험군 분류. 연세대 대학원 석사 학위논문, 2007.

Bastone L, Reilly M, Rader DJ, Foulkes AS. MDR and PRP : a comparison of methods for high-order genotype-phenotype associations. *Human Hered,* 2004:58:82-92.

Breiman L, Freidman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont, CA : Wadsworth Publishing Co., Inc. 1984.

Clark LA, Pregibon D. *Tree-Based Models. Statistical Models in S.* New York, NY: Chapman & Hall, 1993:p377-419.

Cook NR, Zee RY, Ridker PM. Tree- and spline-based association analysis of gene-gene interaction models for ischemic stroke. *Statistics in Medicine,* 2003:23:1439-1453.

Jinbo Chen, Kai Yu, Ann Hsing, Terry M. Therneau. A Parially Linear Tree-Based Regression Model for Assessing Complex Joint Gene-gene and Gene-environment Effects. *Genetic Epidemiology*, 2007:31:238-251.

Loh, W. Y., Shih, Y. S. Split selection methods for classification trees. *Statistica Sinica*, 1997:7;815-40.

Lim, T. S., Loh, W. Y., Shih, Y. S. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning*, 2000:40;203-28.

Marshall, R. J. Partitioning methods for classification and decision making in medicine. *Statistics in Medicine*, 1986:5;517-26.

Marshall, R. J. The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology*, 2001:54;603-9.

Marshall, R. J. *A Manual for SPAN*. 2006.

Shih, Y. S. *QUEST User Manual*. 2004.

Terry M. Therneau, Beth Atkinson. *The rpart Package*. 2008.

Therneau T, Atkinson EG. An introduction to recursive partitioning using RPART routines. *Mayo Foundation technical report*. 1997.

Yu K, Xu J, Rao DC, Province M. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. *Annals of Human Genetics*, 2005:69:577-589.

Zhang HP, Bonney G. Use of classification trees for association studies. *Genetic Epidemiology*, 2000:19:323-332.

ABSTRACT

An application of PLTR to prediction model for the development of liver cirrhosis

Im, Yong-Jin
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

Many statistical methods are used to measure various features that is harmful to diseases. Of which regression analyses and tree-based methodologies are mostly applied. But these methods have some limitations in terms of dealing with multiple variables, because as the number of risk factors increase, their interactive effects to assess are also increased.

As an alternative for these methods, partially linear tree-based regression model(PLTR) is recommended, and which combined the advantages of generalized linear model and tree type model (Jinbo Chen et al., 2006). This method can measure the main effects and interaction effects of risk factors at the same time by means of cooperating linear model with non-parametric tree model.

In this thesis, to evaluate the performance of this method, we applied PLTR method to the real clinical data composed of 4,093 individuals who received the screening test in first and then visited Yonsei University Medical Center for check-up liver cirrohsis from May 1994 to September 2005. For this analysis, we compared a measurement of sensitivity,

specificity, accuracy of all data with logistic regression analysis, QUEST, SPAN.

In the results, we found that the risk factors for liver cirrhosis by PLTR were HBsAg, Anti-HCV, family history, history of drinking, platelet & alkaline phosphate. The PLTR showed the main effects and interaction effects of risk factors simultaneously and that was more accurate for prediction than logistic regression analysis, QUEST. In conclusion, it was confirmed that this PLTR could be appropriately suited to build a prediction model for the development of liver cirrhosis.

Key words : PLTR, Logistic regression, QUEST, SPAN, Sensitivity, Specificity, Accuracy