건강검진 자료에서 성장곡선을 이용한 간 질환 예측모형

연세대학교 대학원
의학전산통계학과
김 영 선

건강검진 자료에서 성장곡선을 이용한 간 질환 예측모형

지도 손 소 영 교수

이 논문을 석사 학위논문으로 제출함

2002년 6월 일

연세대학교 대학원

의학전산통계학과

김 영 선

김영선의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2002년 6월 일

제 목 차 례

그 림 차 례

표 차 례

국 문 요 약

건강검진 자료에서 성장곡선을 이용한 간 질환 예측 모형

본 논문에서는 간 질환 관련 Risk Factor의 연구 및 간 질환 예측 모형을 위해 1994년부터 2001년까지 약 8년간 수집된 건강검진 자료를 토대로 연구하였다. 간과 관련된 기존의 연구에서는 간암에 대한 연구가 주를 이루었지만 간암 발생 전 단계인 간 질환에 대해서는 연구가 거의 이루어지지 않았다. 약 8년간 수집된 55093명의 검진자를 대상으로 간 질환에 관한 Risk Factor와 성장곡선을 통한 예측모형을 추정한 결과 간암에 대한 Risk Factor 대부분이 간 질환에 대해서도 주요한 Risk Factor로 나타나고 있었으며 더 나아가 건강검진 모든 항목들이 Risk Factor가 될 수 있었음을 살펴볼 수 있었다. 또한 건강검진 횟수와 간 질환의 발병률을 살펴본 결과 검진을 많이 받은 사람일수록 발병률이 낮게 나타나고 있음을 볼 수 있었다. 즉 건강에 대해 항시 관심을 가지며 주기적인 검진을 받을 경우 간 질환을 예방할 수 있다는 의미이다. 더 나아가 미래의 시점에서 간 질환 발병에 대해 살펴보고자 다양한 형태의 성장곡선 분석을 통해 Risk Factor를 추정하였으며 이 추정치를 사용한 로지스틱 회귀모형과, 나무모형, 신경망 모형을 통해 살펴보았다. $\hat{X}_{i(5)} = \alpha_i + \beta_i \sqrt{\times} T + \varepsilon_{iT}$의 형태의 성장곡선의 추정치를 사용한 신경망 모형의 경우 간 질환에 대한 accuracy가 72.55% 였으며 Sensitivity는 78.62%로 우수한 모형을 나타내고 있다. 반면 비교 기준이 되는 최근 건강검진 자료의 추정치 (4번째 시점)에서의 간 질환 예측 모형의 경우 accuracy는 72.09% 였으며 Sensitivity는 71.72%로 성장곡선 분석에 의한 간 질환 예측모형에 비해 낮은 수치를 나타내고 있다. 성장곡선과 여러 가지 판별모형에 의해 추정된 다양한 간 질환 예측모형에서 성장곡선 추정값을 사용할 경우 향상된 Sensitivity값을 가져올 수 있었으며 판별모형의 형태 (로지스틱 회귀모형, 나무모형, 신경망모형)에

따라 Accuracy의 차이를 나타내고 있었음을 본 논문을 통해 확인해 볼 수 있었다.

핵심되는 말 : 성장곡선, 건강검진, 로지스틱 회귀분석, 나무모형, 신경망모형
Accuracy, Sensitivity, Specificity, 간 질환, Risk Factor

제 1장 서 론

한국인에게 있어서 간암은 위암 다음으로 많이 발생하는 악성종양(Tumor)으로써 우리 나라 인구 10만명당 20.7명이 간암으로 사망하는데 이것은 세계에서 제일 높은 수치라고 한다(통계청 99년). 현재 조사된 자료에 의하면 남자의 간암 발생률은 10만명당 31.7명으로 여자(10만명당 9.5명, 99년 기준)보다 네 배나 많고 연령이 50대인 경우 남자의 사망률은 여자보다 6.1배나 많다고 한다. 또한 간 질환은 사망원인 순위에 있어서도 1990년 ~ 2000년 사이의 기간에서 5위를 기록함으로써 빈도가 높은 사망원인 중 하나이다. (통계청 2001년 자료) 간 질환은 지난 90년에 비해 사망률이 점차로 감소하고 있는 사망원인이지만 연령별 사망원인의 분포에 있어서 40대를 넘어서는 시점에서 사망원인 순위가 2 위로 갑작스런 상승을 나타내고 있음을 볼 수 있다. 이처럼 간 질환 및 간암은 잠재기를 거쳐 특정 연령 대에 이르면 갑작스런 발병을 하는 위험한 질병이다. 반면 이러한 간 질환에 대한 High Risk Group에 속한 그룹에 대한 조기발견 및 관리가 이루어진다면 사망률을 충분히 줄일 수 있는 질병이라고 생각되어진다.

간암 발생의 주요원인에 대해서 현재 연구된 자료에 의하면 B형과 C형 간염바이러스에 의한 감염과 유전적인 원인 및 가공식품에 첨가된 약품, 음주력과 흡연력, 간기능 검사항목, 알파태아성 단백질(B형 간염)이 간암의 risk factor로 추정되어 진다. 그 외에도 덴마크에서 연구된 보고서에 의하면 산업별 직종에 따라 간암 발생률이 다르게 나타나고 있음을 볼 수 있으며 일본에서 보고된 또 다른 보고서에 의하면 당뇨병이 암 유발에 중요한 요인으로 추정되어지며 특히 간암과 같은 경우 다른 암에 비해 당뇨병환자의 발병률이 높다고 보고되어진다. 간암이 한국인에게 있어서 높은 사망률을 나타내는 질병임에도 불구하고 국내외에서 연구 발표된 논문 대부분은 간암 혹은 간 질환에 대한 Risk Factor의 추정 또는 case-control study에서 간암환자 집단과 정상집단간의 모형의 추정이 대부분이었

다. 그러나 이러한 study에서는 간 질환 또는 간암에 대한 발병의 개별적 원인에 대해 추정만을 할 뿐 간 질환 및 간암의 발병 전의 예후에 대한 언급은 없다. 간 질환에 대한 치료기술의 발달로 인하여 조기발견 및 충분한 치료가 이루어진다면 5년 생존률을 증가시킬 수 있으면 더 나아가 완치도 가능하리라고 생각되어진다.

본 논문에서는 지난 10년간 건강검진(Screening Test)을 받은 검진자를 대상으로 간암 발암 직전 단계인 간 질환에 대한 예후 추정과 간 질환 발생에 대한 예측모형의 추정하는 것이 첫 번째 연구과제이며 두 번째로 미래시점에서 건강검진을 받게 되었을 경우 성장곡선을 통해 간 질환에 대한 Risk Factor 값을 추정하여 간 질환 예측 모형을 추정하는 것이다. 이러한 간 질환에 대한 예측모형은 간암 조기진단 이라는 예방적인 측면에서 많은 도움을 줄 수 있으며 간 질환의 건강검진을 향상시키는 하나의 지표를 마련할 수 있을 거라 생각되어진다.

본 논문에서의 분석방법론은 통계적 방법에 기초한 모형과 데이터 마이닝 방법론에 기초한 모형으로 간 질환에 대한 연구와 성장곡선을 통해 간 질환에 영향을 주는 Risk Factor들의 추정치를 통해 미래 시점에서의 간지질환 예측 모형의 추정에 대해 연구하고자 한다.

제 2장 Screening Test Data

2.1 Screening Test Data (건강검진 자료)

2.1.1 Introduction

건강검진 (Screening Test)는 각종 질병에 대한 위험인자를 찾아내어 질병에 대한 조기진단 및 현재의 건강 상태를 파악하고자 하는 검사이다. 본 논문에서는 간 질환에 대한 Risk Factor 및 예후를 규명하기 위해 건강검진 센터에서 1994년 5월 30일 ~ 2001년 12월 31일 사이 건강검진을 받은 55093명의 검진자를 대상으로 간 질환에 대한 연구를 시작하였다. 건강검진 검사 항목 중 기존의 국내·외에서 간 질환 및 간암에 대한 Risk Factor로 알려진 성별 및 연령 신체계측 지수, 간기능 검사 항목들, 초음파검사, 알파태아성단백, 당뇨병력 및 뇨검사수치, 영양소 섭취량, 음주력, 흡연력과 직업수준에 따라 간 질환에 대한 연구를 시작하였다. 55093명의 건강검진자 중 건강검진을 2번 이상을 받은 검진자는 9919(18%)였으며 많게는 9번에 걸쳐 건강검진을 받기도 하였다. 한 가지 주의 깊게 살펴봐야 할 사항은 건강검진을 많이 받은 사람일수록 간 질환에 대해 양호한 결과를 나타내고 있는데 이는 평상시 건강에 대해 관심을 가질수록 건강하다고 추정되어진다. 따라서 기존의 Study에서 알려진 간 질환에 대한 Risk Factor에 건강검진 횟수 역시 중요한 변수로 추가되어 져야 한다. 이러한 건강검진자료(Screening Data)는 간암에 대한 연구에서 간암 여부를 알 수 있는 정밀검진이 실시되지 않기 때문에 간암 예측모형에서의 종속변수(간암여부)에 대한 정보가 존재하지 않는다. 반면 간 질환 여부(종속변수)는 초음파 검사를 통해 알 수 있다. 초음파 검사 결과 지방간이나 간 종양, 간경화와 같은 간 질환을 정확히 찾아낼 수 있으며 이러한 검사 결과는 간 질환에 대한 Risk Factor 및 예후를 추정하는 모형에서 중요한 종속변수가 되어질 수 있다.

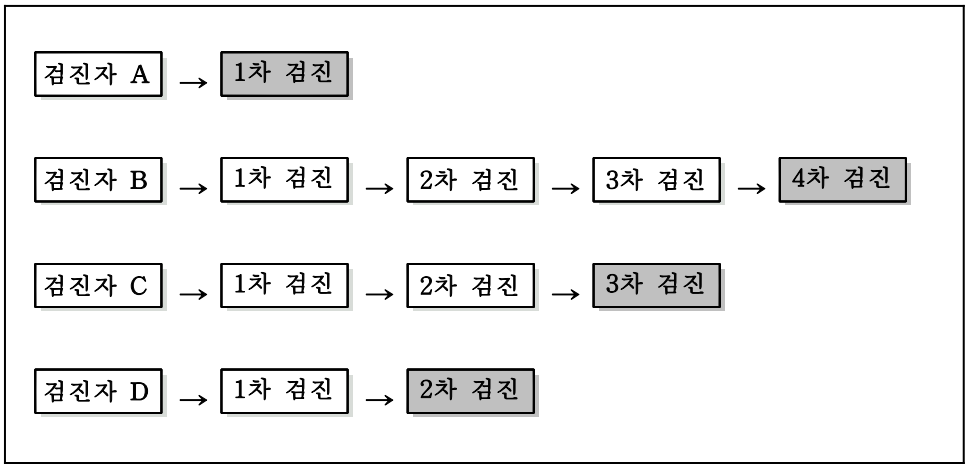
초음파 결과를 통해 간 질환과 관련되었다고 추정할 수 있는 종속변수의 설정은 연세대 소화기 내과 선생님들을 통해 정의되어졌다. 이러한 정보에 기초하여 간 질환에 대한 설명력 높은 원인 규명 및 예측 모형을 제시할 수 있다면 간 질환 환자들의 간암으로의 발병률을 낮출 수 있으리라고 생각되어진다.

2.1.2 DATA I. 전체 자료

본 논문에서 연구된 건강 검진자 55093명의 검진 횟수는 많게는 9번에서 1번까지 있다. 이러한 자료의 형태에서 검진자의 검진 횟수를 모두 다른 관찰치로 사용하는 분석 자료를 만들었다. 즉 A라는 검진자가 4번의 검진을 받았으면 분석 모형에서는 4개의 관찰치가 들어가게 된다. 전체 자료의 분석은 검강검진 횟수와 간 질환의 발병률간의 연관성이 있는지에 대해 살펴보기 위한 중요한 자료 형태라고 생각되어진다.

DATA II. 최근 자료 (Discriminant Analysis Data)

그림 1 간질환 판별 모형 추정을 위한 자료

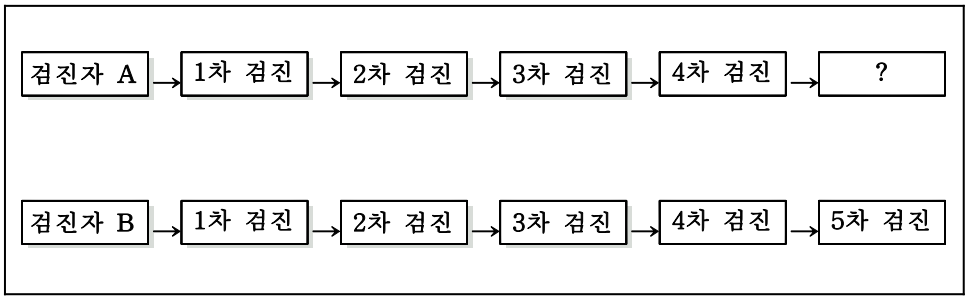


검진자가 여러 번 검진을 받았다고 하더라고 가장 최근의 검진 결과만을 관측치로 사용하여 분석 자료를 만들었다. 그림 1에서 나타나듯이 B라는 검진자가 4번

의 검진을 받았다고 하더라도 가장 최근 받은 검진 결과만이 분석 자료에 포함이 되고 전에 받은 3번의 검진 결과는 분석 대상에서 제외된다. 이러한 자료를 사용하는 이유는 건강 검진 횟수에 따라 간 질환 발병률이 다르게 나타나고 있기 때문이다. 검진을 단 한 번만 받았던 사람들의 간 질환 발병률이 9.45%인데 반해 4번의 검진을 받았던 사람들은 간 질환 발병률이 1.28% 였다. 즉 검진을 많이 받은 사람들의 과거 검진 자료 모두를 분석대상으로 한다면 전체 자료에서 간 질환 발병률이 낮아지는 bias가 발생하기 때문이다.

2.1.3 DATA II. Predictive Data (Growth Curve Analysis)

그림 2 성장곡선을 추정후 간 질환 예측 모형 자료



성장곡선분석을 통해 간 질환 영향을 주는 Risk Factor의 앞으로의 변화량에 대해 추정 후 이러한 추정값을 통해 미래시점에서의 간 질환 예측모형을 추정해 볼 수 있다. 이러한 성장곡선 분석을 위해 사용된 검진자는 최소 4번 이상의 검진을 받은 사람들을 대상으로 하였다. 이렇게 4번의 검진을 받은 사람은 1532명이었으며 실제 분석 가능한 자료는 1030명으로 간 질환 판별 모형 분석에 있어서 Training Set으로 사용되어질 수 있으며 이 1030명의 검진자 중에서 실제 5번 이상의 건강 검진을 받은 사람은 678명 있었으며 실제 분석 가능한 자료는 430명이었다. 이렇게 실제 5번의 검진을 받은 검진자를 간 질환 예측 모형의 Validation Set으로 사용할 수 있다. 5번째 검사를 받은 검진자는 실제 5번째 시점에서 간 질

환이 있는지 없는지에 대한 초음파 정보를 가지고 있기 때문에 1030명을 Training Set으로 사용하여 추정한 간 질환 예측모형에서 나온 결과와 실제 결과를 비교해 볼 수 있는 좋은 자료라고 생각되어 진다.

2.2 Description of Screening Test Data Variable

현재 건강검진 센터에서 검진하고 있는 항목들을 살펴보면 기초정보, 신체계수, 폐기능, 혈액검사, 혈청검사, 생화학검사, 영양소 섭취정도, 음주 및 흡연에 대한 과거력, 초음파(복부) 등이 있다. 이러한 자료 중 간 질환에 기존에 알려진 Risk Factor는 기초정보, 신체계수, 생화학검사, 과거력 등이다. 이러한 검사항목들이 과연 간 질환에 어느 정도 위험인자로 존재하며 더 나아가 간암 발현에 어느 정도 영향을 미치는지 살펴볼 수 있다. 세브란스 건강검진 센터에의 검사되어지는 세부 항목은 표1과 같다.

표 1 건강검진 검사 항목

검사항목	하위항목	변수	검사항목	하위항목	변수
기초정보	기초정보	성별	생화학검사	대사 및 전해질	나트륨
		연령			칼륨
		건강검진횟수			염소
신체계수	신체계수	BMI			이산화탄소
		신장			칼슘
		체중			인
		표준체중			혈당
폐기능	폐기능	노력성폐활량			혈중요소질소
		최고호기유속			크레아티닌
		1초간노력상호기량			요산
		1초간호기량/폐활량비		간기능	총단백
혈액검사	적혈구	RBC			알부민
		Hb			총빌리루빈
		Hct			Alk.Phos
		MCV			Ast(GOT)
		MCH			Alt(GPT)
		MCHC			γ-GT
	백혈구	임파구			LDH
		호산구		혈청지질	총콜레스트롤
	혈소판	혈소판			중성지방
		ABO			고밀도콜레스트롤
		Rh		갑상선기능	T4
영양소 섭취량	영양소섭취량	철분			T3
		나이아신			갑상선자극호르몬
		열량		종양혈청	전립선특이항원(남)
		단백질		뇨검사	비중
		지방			산도
		당질			단백
		칼슘			요당
		인			케톤체
		비타민A			잠혈
		비타민B1			요빌리노겐
		비타민B2			빌리루빈
		비타민C			아질산염
혈청검사	혈청검사	c반응성단백			백혈구
		류마티스인자			탁도
					색

제 3장 연구모형

3. 1. 간암 및 간 질환에 대한 Risk Factor

3.1.1 B형과 C형 간염바이러스

현재 보고된 자료들에 의하면 간암에 결정적인 영향을 미치는 Risk Factor로 B형과 C형 간염바이러스가 있다. 이러한 B형 및 C형 간염 바이러스는 지역간의 차이가 상당히 많이 존재한다. 『간암의 빈도가 높은 동북아시아 및 아프리카 지역에서 간염에 대한 감염률이 높다고 보고되어지는 것』은 이를 잘 증명해 주는 내용이다. 2001년에 발표된 『International Trends and Patterns of Primary Liver Cancer』에 의하면 Thailand의 Khon Kend을 비롯한 아시아 지역의 간암 발생률이 가장 높게 나타나고 있음이 보고되어졌으며 Oceania와 South and Central America의 지역에서의 발생률이 가장 낮게 나타나고 있다고 한다. 본 논문에서는 간암에 대한 분석은 아니며 이러한 B형과 C형 간염 바이러스는 하나의 종속변수로 사용되어진다.

3.1.2 음주력과 흡연력

술을 많이 마시거나 담배 흡연이 심한 경우 모든 질병 발생의 중요한 원인이 된다. 음주력과 흡연력과 같은 과거력은 간 질환 및 간암에 중요한 Risk Factor로 알려져 있으며 음주와 흡연은 간기능에 직접적인 영향을 미치는 것으로 보고되어져 있다. 이러한 음주 및 흡연은 주로 여성보다는 남성이 심각한 수준이며 20세를 전후로 시작이 이루어진다. 음주 및 흡연 당시는 영향을 많이 미치지는 못하지만 10년 이상 지속적으로 이루어 졌을 경우 모든 질병을 일으키는 중요한 요인이 된다. 그러나 건강검진 센터에서 음주력 및 흡연력에 대한 정보수집을 최근부터 시작되었기 때문에 자료의 분석시 결측치의 비율이 전체 자료의 90%이상이 된다.

그러므로 본 분석에서는 음주력과 흡연력에 대한 정보를 제외하였다. 그러나 음주와 흡연으로 인한 폐 기능 수치와 간 기능 수치로 대체하여 분석하고자 한다.

3.1.3 성별

지금까지 보고된 간암 환자에 대해 조사하여 보면 여성보다는 남성의 간암 발병률이 높은 것으로 알려져 있다. 성염색체상에 간암 발생에 영향을 미치는 유전자가 존재하는지에 대해 아직 보고된 바는 없으나 간암은 부모로부터 유전되어 발생되어진다는 연구결과는 있었다. 하지만 간암이 성염색체상 존재 여부에 대해 아직 판명되지는 않았지만 일반적으로 남성이 여성에 비해 음주 및 흡연을 많이 하며 남성이 여성에 비해 사회활동을 하는 경제인구비가 높다. 이러한 영향으로 인해 성별간에 간암발생의 차가 나타나고 있으며 이는 간암 발생의 중요한 Risk Factor로 간주되어 진다.

3.1.4 연령

간암 발생에 대해 연령이 많은 영향을 주는 것으로 보고되어지고 있다. 5세 이전에는 간암 발생률이 아주 낮으며, 성별간 차도 없지만 이 시점을 이후부터 서서히 증가 추세를 나타내며 남자의 경우 30세부터 49세 사이에 발생률이 현저하게 높게 나타나고 있으며 그 후에도 계속 증가하여 연령이 5년씩 높아질수록 간암 발생률은 약 3배씩 증가하여 70~74세에서는 인구 10만명당 290명정도로 높은 발생률을 나타내고 있다.

3.1.5 당뇨병력

일본 lukuoka에 거주하는 30세에서 79세 사이의 주민을 대상으로 1986년부터

1989년까지 연구되어진 자료를 토대로 간암 및 암들에 대해 성별과 연령, 음주력과 흡연력 그리고 당뇨병에 대한 병력이 간암 발생에 영향을 미친다고 보고되어졌다. 당뇨병력이 있는 사람들에 대해 연령과 성별을 adjusted하여 relative risk를 보면 1.59 (95%C.I. =1.14 ~ 2.23; p=0.007)로 당뇨병력이 없는 사람에 비해 59%가 더 높은 암 발생률을 나타내고 있다. 또한 음주력과 흡연력을 adjusted하여 암에 대한 relative risk를 보면 1.57 (95%C.I = 1.12-2.20;p=0.008)로 높게 나타나고 있다. 간암에 대해 살펴보면 연령과 성별을 adjusted하여 relative risk가 2.82 (95%C.I. =1.58 ~ 5.03; p=0.001)이며 음주력과 흡연력을 adjusted하면 간암에 대한 당뇨병력의 relative risk는 2.75 (95%C.I.=1.54 ~ 4.91)로 다른 암에 비해 높은 관련성을 나타내고 있다.

3.1.6 종사직종

덴마크에서 연구된 결과에 의하면 산업별 종사직종에 따라 간암의 발생률 차이가 난다고 보고되어졌다. 각 직종에 대해 근무연수 10년 이상인 집단과 10년 미만인 집단에 대해 각각의 Odd Ratio를 비교하였을 때 위생직이나 자동차 정비, 술 제조장, 제지공장에서 10년 이상 근무하였을 경우 간암의 Odd Ratio가 높게 나타나고 있음을 살펴볼 수 있다. 또한 이러한 직종에 근무한 10년 미만의 종사자들에 비해서도 Odd Ratio가 2~4 정도의 차이를 나타내고 있음을 『Liver Cancer Among Employees in Denmark』을 통해 볼 수 있었다.

3.1.7. 간기능 수치

과거 간기능 수치를 통해 간 질환 여부를 판단하였지만 정확도가 떨어져서 현재는 간 질환에 대한 판단 기준으로는 사용되어지고 있지는 않지만 간 상태에 대한 중요한 정보를 제공하고 있다.

3.2. 간 질환 판별모형

3.2.1 간질환 예측모형

먼저 간암예측모형에서 대해 몇 가지 고찰하여야 하는 부분에 대해 생각하여 볼 수 있는데 첫 번째로 간 질환에 영향을 주는 Risk Factor의 정도에 따른 연구가 필요 되어지며 이러한 Risk Factor에 의해 추정된 간 질환 예측모형의 우수성에 대해 평가해 볼 수 있다. '간암양성자'를 '간암음성자'에 대해 event가 발생한다고 가정을 하였을 경우 예측모형으로 추정한 결과 Classification Table를 작성하여 모형의 우수성에 대해 비교할 수 있는데 Classification Table에는 각 관측치들에 대해 모형으로부터 예측된 확률값이 나오는데 이 확률값을 통해 '간암양성자'와 '간암음성자'를 분류하여 낸다. 어느 그룹에 속하는지에 대한 분류의 기준이 되는 예측 확률값에 대해 먼저 『로지스틱 모형』,『나무 모형』,『신경망 모형』에 대해 이론적인 고찰을 통해 생각해 볼 수 있다.

식1. 로지스틱 회귀분석에 의한 간질환 예측모형

$$\theta_h = \{1 + \exp[-\alpha - \sum \beta_k x_{hx}]\}^{-1}$$

식2. 신경망 분석에 의한 간 질환 예측모형

$$y_k = g'\left(\sum_j^M w_{kj}^{(2)} g\left(\sum_{i=0}^d w_{ji}^{(1)} x_i\right)\right)$$

g, g' :activity function

w : weight

먼저 식1에 나타난 로지스틱 모형과 식2에 나타난 신경망 모형, 그리고 나무모

형과 같은 판별 모형들을 통하여 간 질환에 대한 예측모형의 추정 및 간 질환에 대한 예측 확률값을 구할 수 있다. 이러한 확률값에 대해 일반적으로 default로 설정된 cut-off value가 0.5는 종속변수 2개의 집단의 sample size가 동일하거나 거의 비슷한 경우 유용한 기준으로 사용되어질 수 있으나 2개 집단의 sample size가 unbalance한 경우에는 size가 작은 쪽에 대해 왜곡되는 현상이 일어난다. 즉 두 집단의 사전분포(Prior Probability)가 존재하게 되며 이를 수정하기 위해서는 적절한 cut-off value를 찾아서 'Sensitivity'와 'Specificity', 'accuracy'를 향상시키는 적절한 수준을 찾아내야 한다. 이러한 관계에 대해 표2를 보고 살펴보자.

표 2 Confusion Matrix

		From Predict Model	
		간암음성자	간암양성자
DATA	간암음성자	Specificity	False Positive
	간암양성자	False Negative	Sensitivity

자료에서 나타난 결과와 모형에 의해 추정된 결과에 대해 위의 표와 같은 Confusion Matrix를 작성할 수 있다. 자료에서 '간암양성자'며 또한 예측된 모형에서 '간암양성자'로 판별된 경우를 일반적으로 'Sensitivity'라고 할 수 있다. 또한 자료에서 '간암음성자'를 예측모형으로 판별된 경우를 'Specificity'라고 한다. Accuracy는 'Sensitivity'와 'Specificity'에 영향을 받아서 두 값들이 모두 높아야 Accuracy를 높일 수 있다. 이러한 Accuracy는 판별 모형의 경우 Predict Value의 값이 0.5일 때 가장 높게 나타난다. 반면 높은 Accuracy 값을 가진다고 하여도 적절한 수준의 Sensitivity와 Specificity의 값을 가지기 어려운데 이러한 경우 분석자에 의해 그 적절 수준을 결정하는 것이 필요 되어진다고 생각된다.

제 4장 간 질환 Risk Factor 및 판별 모형 결과

4. 1. 검사횟수

건강검진 센터에서 1994년 5월 30일 ~ 2001년 12월 31일 사이 건강검진을 받은 55093명의 검진자를 대상으로 간 질환에 대한 분석을 실시하였다. 또한 55093명의 건강검진자 중 건강검진을 2번 이상을 받은 검진자는 9949명 (13.92%)이었으며 3번 이상 검진을 받은 검진자는 3793명(5.31%)이었다. 한 가지 주의 깊게 살펴봐야 할 사항은 건강검진을 많이 받은 사람일수록 간 질환에 대해 양호한 결과를 나타내고 있는데 이는 평상시 건강에 대해 관심을 가질수록 건강하다고 추정되어진다. 건강검진을 받은 횟수에 따라 간 질환 발병률을 살펴보면 건강검진을 많이 받으며 건강에 관심을 가지고 있는 사람일수록 그 비율이 낮아짐을 볼 수 있다.

표3. 건강검진 횟수에 따른 간 질환 발병률

	TOTAL				NEGATIVE				POSITIVE				
TIME	N	%	누적 N	누적 %	N_1	%	누적 N	누적 %	N_2	%	누적 N	누적 %	RATE
1	45174	82.00	45174	82.00	43522	81.77	43522	81.77	1652	88.58	1652	88.58	3.66
2	6132	11.13	51306	93.13	5978	11.23	49500	93	154	8.26	1806	96.84	2.51
3	2255	4.09	53561	97.22	2209	4.15	51709	97.15	46	2.47	1852	99.3	2.04
4	854	1.55	54415	98.77	842	1.58	52551	98.73	12	0.64	1864	99.95	1.41
5	376	0.68	54791	99.45	375	0.7	52926	99.43	1	0.05	1865	100	0.27
6	213	0.39	55004	99.84	213	0.4	53139	99.83					
7	68	0.12	55072	99.96	68	0.13	53207	99.96					
8	20	0.04	55092	100.00	20	0.04	53227	100					
9	1	0.00	55093	100.00	1	0	53228	100					
TOTAL	55093				53228				1865				3.38

이러한 관계에 대해 통계적으로 유의한지 여부를 알아보기 위해 chi-square Test를 한 결과 '건강검진을 많이 받은 사람일수록 간 질환의 발병률이 낮아진다' 라고 추정되어진다. (chi-square 검정통계량= 67.13, 자유도= 3, p-value=0.0001)

즉 건강에 대해 항상 관심을 가지고 주기적인 건강검진과 그 결과에 따라 자신의 생활을 조절하는 사람들은 질병의 발병률이 낮아짐을 본 분석을 통해 살펴볼 수 있었다.

4.2 간 질환 판별 모형의 추정

2번 이상 건강검진을 받은 검진자 9919명 중 Data Cleaning을 통해 사용 가능한 7806명을 대상으로 하여 간 질환 판별 모형을 연구하였다. 먼저 가장 최근 전에 건강검진을 받은 자료를 Training Set으로 사용하여 모형을 추정하였고 가장 최근에 건강검진을 받은 자료를 Validation Set으로 사용하여 모형에 대한 평가를 실시하였다. 판별 모형에 대한 추정 방법은 로지스틱 회귀분석, 나무모형, 신경망 모형들을 통해 간 질환에 영향을 주는 Risk Factor에 대한 살펴보았다.

4.2.1. Logistic Regression을 통한 간 질환 판별 모형

표4. 로지스틱 회귀분석을 통한 간질환 판별 모형

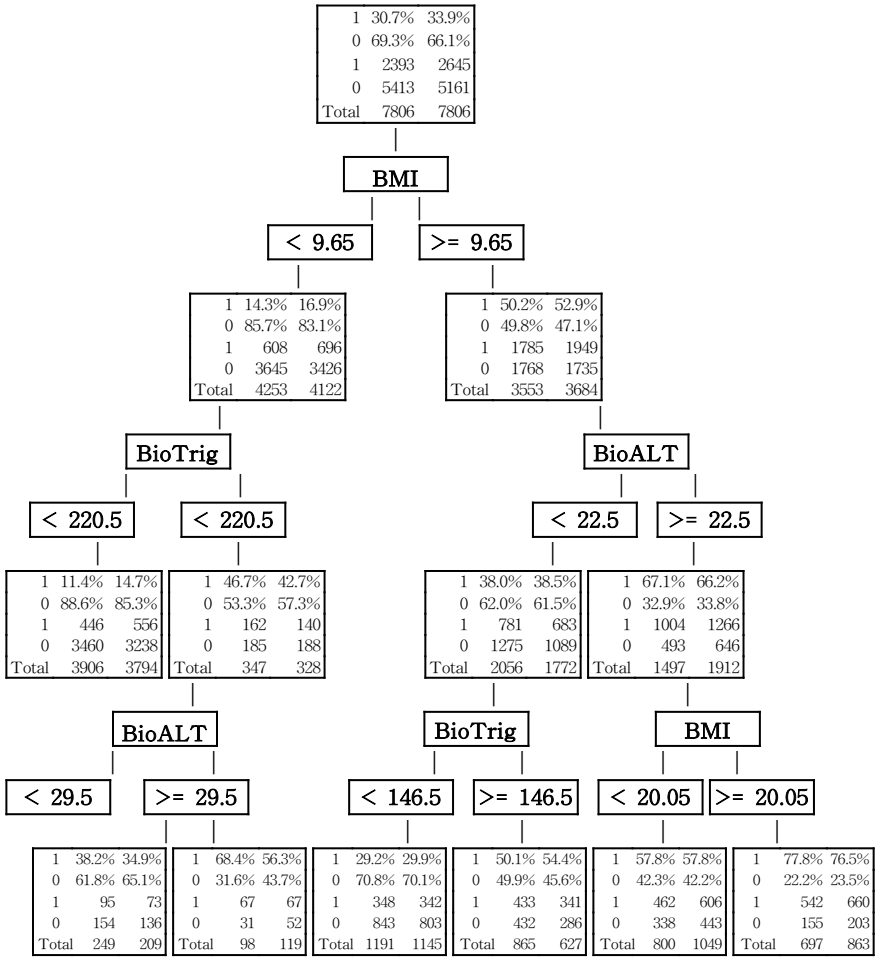
Parmeter	Estimate	SE	Wald chi-square	95% C.I. Lower	95% C.I. upper	p-value
Intercept	-5.3809	0.5689	89.46	-6.4959	-4.2658	0.0001
BMI	0.0692	0.00255	735.57	0.0642	0.0742	0.0001
BioAlb	0.4080	0.0965	17.87	0.2188	0.5972	0.0001
BioAlt	0.0169	0.00167	101.43	0.0136	0.0201	0.0001
BioChol	0.00432	0.000853	25.66	0.00265	0.00600	0.0001
BioCO2	-0.0345	0.0114	9.20	-0.0568	-0.0122	0.0024
BioGlucose	0.00998	0.00130	59.32	0.00744	0.0125	0.0001
BioHdl	-0.0163	0.00279	34.19	-0.0218	-0.0108	0.0001
BioT4	-0.0231	0.00835	7.64	-0.0395	-0.00671	0.0057
BioTrig	0.00293	0.000340	74.00	0.00226	0.00359	0.0001
BioUric	0.0833	0.0249	11.21	0.0345	0.1320	0.0008
age	0.0237	0.00287	68.59	0.0181	0.0294	0.0001
Fvc	-0.00827	0.00198	17.46	-0.0121	-0.00439	0.0001
sex	0.1537	0.0370	18.05	0.0847	0.2299	0.0001

먼저 bioAst와 bioALT의 상관력이 약 -77%로 높게 나타나고 있어 다중공산

선의 문제 때문에 bioAst를 제외하고 로지스틱 회귀분석한 결과이다. 간 질환에 가장 영향을 많이 주는 Risk Factor로는 간 기능 수치인 BioAlb였으며 그 외에도 간 기능 수치들과 신체지수인 BMI, 혈청지질 수치와, 뇨검사 수치들 등이 간 질환에 영향을 주는 주요 Risk Factor로 추정되어졌다. 이러한 Risk Factor 들은 간암 관련 논문들에서 이미 언급된 것들로 간 질환 역시 간암과 비슷한 수준의 Risk Factor를 가지고 있는 것으로 밝혀졌다.

4.2.2 간 질환 판단을 위해 나무 모형

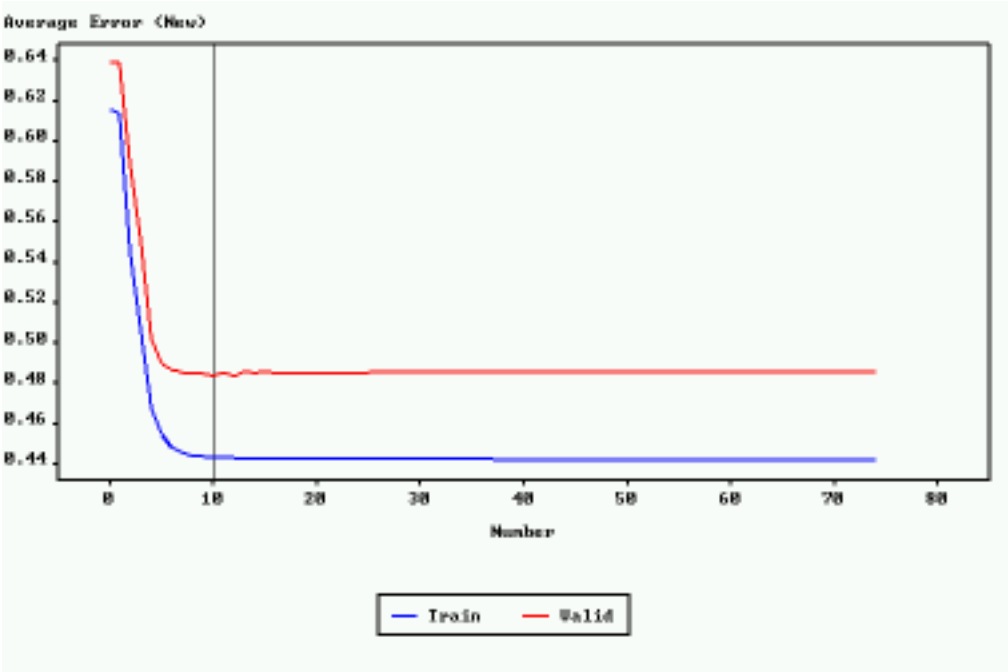
그림3. 나무모형을 이용한 간 질환 판별 모형



나무 모형을 보면 먼저 BMI, BioTrig, BioAlt 3개의 Risk Factor 수준에 따라 간 질환 판별 모형을 추정해 볼 수 있다. 먼저 BMI 수치가 9.65 보다 작고 BioTrig 수치가 220.5 보다 작으면 간질환에 대해 건강하다고 추정해 볼 수 있으며 반면 BMI 수치가 6.65보다 크고 BioALT 수치 역시 22.5보다 크며 BMI 수치가 다시 20.5보다 큰 경우는 간 질환에 위험 군에 속하는 것으로 나무모형을 통해 살펴볼 수 있다. 이 외 5 개의 terminal node에 대해서도 이와 같이 해석해 볼 수 있다.

4.2.3 간 질환 판단을 위해 신경망 모형

그림 4 신경망 모형을 이용한 간 질환 판별 모형시 Average Error Plot



신경망에 의한 간 질환 판별 모형에서 Validation Set의 평균오차가 iteration이 10일 경우 최소를 나타나고 있다. 즉, 10번의 반복과정을 통해 계수값의 개선

(Count Method : $X_n = X_{n-1} + a$)을 하였으며 평균 오차가 가장 적은 경우에서 가장 이상적인 모형을 추정하였다. 신경망 모형의 목적함수(Object Function)의 추정치들에 대해서는 생략하였다.

4.3 간 질환 판별 모형 평가

표5. 간질환 판별을 위한 Validation Set Classification Table

Model	predicted probability	specificity	sencitivity	false negative	false positive	Total Accuracy	Miss clasification
LOGISTIC	0.1	34.97	95.20	65.03	4.80	55.38	44.62
	0.2	55.86	89.26	44.14	10.74	67.18	32.82
	0.3	69.70	80.34	30.30	19.66	73.30	26.70
	0.4	80.39	68.88	19.71	31.12	76.43	23.57
	0.5	87.19	56.75	12.81	43.25	76.88	23.12
	0.6	92.09	43.71	7.91	56.29	75.70	24.30
	0.7	95.74	31.15	4.26	68.85	73.86	26.14
	0.8	97.69	19.58	2.31	80.42	71.23	28.77
	0.9	99.19	8.62	0.81	91.38	68.50	31.50
	1.0	100.00	0.00	0.00	100.00	66.12	33.88
TREE	0.1	0.00	100.00	100.00	0.00	33.88	66.12
	0.2	62.74	78.98	37.26	21.02	68.24	31.76
	0.3	78.30	66.05	21.70	33.95	74.15	25.85
	0.4	86.57	55.12	13.43	44.88	75.92	24.08
	0.5	86.57	55.12	13.43	44.88	75.92	24.08
	0.6	90.95	43.25	9.05	56.75	74.79	25.21
	0.7	96.07	24.95	3.93	75.05	71.98	28.02
	0.8	100.00	0.00	0.00	100.00	66.12	33.88
	0.9	100.00	0.00	0.00	100.00	66.12	33.88
	1.0	100.00	0.00	0.00	100.00	66.12	33.88
NEURAL	0.1	39.66	94.44	60.34	5.56	58.22	41.78
	0.2	55.98	88.51	44.02	11.49	67.00	33.00
	0.3	65.76	82.27	34.24	17.73	71.36	28.64
	0.4	76.67	70.06	23.33	29.94	74.43	25.57
	0.5	83.18	60.30	16.82	39.70	75.43	25.57
	0.6	89.01	49.19	10.99	50.81	75.52	24.75
	0.7	94.56	33.76	5.44	66.24	73.96	26.04
	0.8	98.95	8.39	1.05	91.61	68.26	31.74
	0.9	100.00	0.00	0.00	100.00	66.12	33.88
	1.0	100.00	0.00	0.00	100.00	66.12	33.88

간 질환 판별 모형들의 우수성을 평가해 보면 위의 표에서 나타난 결과와 같이 Cut-off Value 0.3 수준에서 로지스틱 회귀모형에서 가장 높은 Accuracy (76.88%)의 값을 가지고 있지만 간 질환을 간 질환이라고 맞추는 정도인 Sensitivity의 경우 56.75%로 적합한 모형의 형태라고 보기는 어렵다. 따라서 Cut-off Value를 0.3수준으로 낮추었을 때 나무모형의 경우 가장 높은 Accuracy 74.15%를 나타내고 있지만 Sensitivity의 경우 66.05%로 상당히 낮은 수치를 나타내고 있다. 반면 Cut-off Value를 0.3수준에서 신경망 모형의 경우 accuracy가 71.36%로 로지스틱 회귀모형의 accuracy 73.30에 비해 약 2% 정도 떨어진 값을 나타내고 있지만 Sensitivity의 경우 82.27%로 로지스틱 회귀모형의 Sensitivity 80.34에 비해 약 2%정도 높은 값을 가지고 있다. 간 질환 판별을 위한 적절한 모형의 선택과 Cut-off value 수주에 있어서 cut-off value 0.3 수준에서 로지스틱 회귀모형과 신경망 모형이 우수한 모형으로 나타나고 있으며 간 질환에 대한 기회비용을 고려해 보았을 경우 신경망 모형의 선택이 적절하다고 생각되어진다.

제 5 장 성장곡선을 통한 판별모형

건강검진 최근 검사를 통해 간 질환에 영향을 주는 주요 Risk Factor로 추정된 BMI, BioAlb, BioAlt, BioChol, BioCO2, BioGlucose, BioHdl, BioT4, BioTrig, BioUric, age, Fvc, Sex와 같은 건강검진 항목들에 대해 미래시점에서의 이러한 검진 항목들의 수치들을 추정해 봄으로써 시점에서의 간 질환 예측모형을 추정해 보고자 한다. 건강검진 자료중 4번 이상 검진을 받은 검진자의 수는 1030명으로 이들을 대상으로 5번째 시점에 건강검진을 받는다는 가정 하에서 Risk Factor의 변화와 추정치들을 구해 볼 수 있다. 이렇게 추정된 건강검진 항목을 통해 5번째 시점에서의 간 질환 예측모형을 추정해 볼 수 있다.

5.1 성장곡선의 소개 (Introduction of Growth Curve)

4번 이상의 건강검진을 받은 검진자들에 대해 1차 선형모형, 2차 항이 존재하는 선형모형, Log 모형, root 모형등을 통해서 다양하게 fitting 시켜 볼 수 있다. 이렇게 추정된 모형에서 기울기를 나타내는 Parameter가 유의적인 모형들만을 골라 미래 시점에서 건강 검진 항목의 수치가 어떻게 변하는지에 대해 살펴 볼 수가 있다. 반면 기울기를 나타내는 Parameter 유의적이지 않다면 이러한 성장곡선들은 거의 변화를 나타내지 않고 있으므로 4번째 시점에서의 건강검진 항목 수치로 미래시점(5번째)에서의 건강검진항목 수치를 대체하여도 무관하다고 생각되어진다.

그림5. 간기능 수치와 건강검진과의 Plot

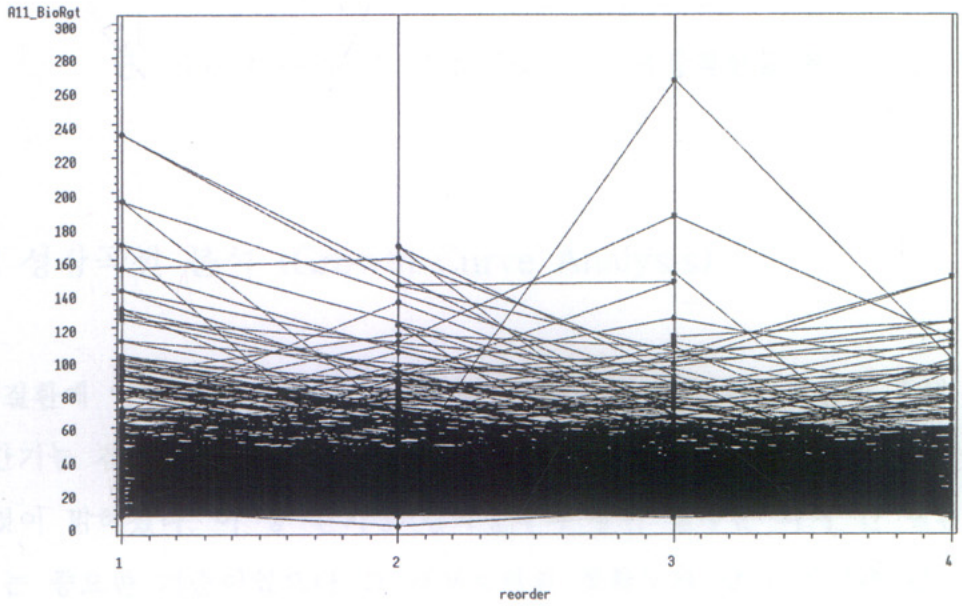


그림5는 4년 동안 건강검진을 받은 사람들을 대상으로 간 기능 수치중 하나인 BioRgt 수치의 변화를 나타낸 것이다. 검진자 전체가 하나의 그래프 안에 overlay 된 모습이 약간은 복잡하게 나타나 있지만 이 그래프를 통해 얻을 수 있는 정보가 있다. BioRgt의 수치가 8이상 50이하의 수준을 가진 검진자들은 그 시간이 흐름에 따라 수치가 거의 변하지 않고 평행하게 나가고 있음을 볼 수 있으며 반면 50이상의 수치를 가진 검진자들은 그 수치가 상승 또는 하락의 형태를 나타나고 있음을 추정해 볼 수 있다. 즉 그래프를 통해 2개의 집단을 추정해 볼 수 있는데 하나가 BioRgt의 수치가 변화가 없는 집단이고 다른 하나는 BioRgt 수치의 변화가 있는 집단이다. 이러한 두 집단에 대해 만약 미래시점에서 간 질환 예측모형의 추정에 있어서 성장곡선 분석을 통해 Risk Factor의 변화가 없는 변수에 대해서는 가장 최근 검진을 받았던 4번째 시점에서의 검사항목을 예측모형에 넣어 분석을 할 수 있으며 반면 성장곡선을 통해 Risk Factor의 변화가 확연히 나타나는 검진자에 대해서는 성장곡선을 통해 추정된 값으로 5번째 검사 시점에서의 간 질환 예측모형의 분석 대상이 될 수 있다.

식3. Risk Factor 추정치

$$\hat{X}_{(5)} = X_{(4)} \text{ or } f(x)$$

단, $f(x)$는 유의적인 기울기를 가진 성장곡선을 통해 얻은 추정치의

5.2 성장곡선 분석 (Growth Curve Analysis)

간 질환에 영향을 주는 Risk Factor에 대해 로지스틱 회귀모형을 통해 BMI 수치와 간기능 검사항목 중 ALT, ALB 그리고 뇨검사 항목 폐기능 수치 등이 유의적인 것이 밝혀졌다. 이 중 간기능 검사항목과 같은 경우는 과거 간 질환 여부를 판단하는 중요한 기준이었으나 그 자체로써의 정확도가 낮아 지금은 참고적인 자료로 활용되어진다. 반면 현재 간 질환 여부를 판단할 수 있는 기준이 되는 초음파 검사와 같은 경우는 정확도는 높으나 그 자체로써 차후 건강증진을 위한 방향성을 제시하기 어려운 변수이다. 예를 들어 간 기능 검사 항목 중 ALT와 같은 수치는 건강한 성인의 경우 8 ~ 30 정도의 수준이며 이 이상의 수치가 나왔을 때는 수치를 낮추는 방향으로 치료와 식이요법 등이 병행되어 질 수 있으나 초음파 검사는 그 자체로써 간 질환 여부를 판별 할 뿐이지 차후 행동을 요구하지 못한다. 이러한 점을 감안하여 앞으로 간 질환에 영향을 주는 Risk Factor에 대한 예측은 간 질환 예방에 도움이 될 뿐 아니라 앞으로 간 질환에 걸리는 확률을 예측함으로써 조기 진단에 많은 도움을 줄 것이라고 기대되어진다. 성장곡선의 분석을 위해 4번 이상의 검진을 받은 1030명을 대상으로 하여 판별 모형을 통해 Risk Factor로 알려진 변수 전체에 대해 1차 모형과 2차 모형 Log 모형 root 모형에 적합을 시켜 보았다. 이러한 적합은 검진자 개별적으로 4개의 시점에 대해 Risk Factor의 변화를 선형과 곡선 모형에 적합 시킨 것으로 차후 검진자가 5번째 혹은 그 이상의 방문시 Risk Factor를 예측하기 위함이다.

5.2.1 일차・이차 성장곡선 모형

먼저 어떤 검진자가 4번의 검진을 통해 간 질환에 대한 Risk Factor들의 값을 얻을 수 있으며 이 값을 기초로 하여 일차선형 모형에 적합 시켜 Risk Factor가 검진 회수에 따라 변동되는 것을 예측할 수 있다. 즉 $X_{i(T)}$를 i번째 검진자가 T시점에서의 ALT 수치라고 가정할 수 있으며 이러한 일차 선형 모형을 다음과 같은 공식으로 생각해 볼 수 있다.

식4. 일차 성장곡선

$$X_{iT} = \alpha_i + \beta_i \times T + \varepsilon_{iT}$$

i : 간기능 검사를 받은 검진자

T: 간기능 검사시점

이러한 일차선형 모형에 대해 적합 시켰을 때 그 모형이 적절치 못할 경우 2차 또는 Log, root 모형에 적합 시켜 볼 수 있다. 반면 4개의 시점에서 3차 방정식에 대한 적합은 과적합(overfitting)되는 경우이므로 2차 또는 Log, root 모형에 대한 적합을 통해 5번째 시점에서의 값을 추정해 볼 수 있다.

식5. 이차 성장곡선

$$X_{iT} = \alpha_i + \beta_{1i} \times T + \beta_{2i} \times T^2 + \varepsilon_{iT}$$

검진자 개개인에 대해 Risk Factor들을 4개의 시점에 적합 시켜 최고차항의 coefficient가 유의적인지 살펴 볼 수 있다. 먼저 유의수준 10%하에서 최고차항의 coefficient가 유의적인 방정식의 수는 1차와 2차에 방정식에 따라 다르게 나타나

고 있다. 1차 방정식의 경우 대체적으로 전체 Risk Factor들 중 평균 14.67%가 유의적으로 나타나고 있으며 반면 2차 방정식의 경우는 10.72%로 1차 방정식 보다 적합하지 못한 것으로 추정되어진다. 대체적으로 간 질환 여부를 추정할 수 있는 Risk Factor들이 파장을 이루며 변화되지만 증가나 감소의 방향성을 가지며 변동되어지거나 변화의 정도가 거의 없는 두 개의 그룹으로 나뉘어지는 것을 알 수 있다.

표6. 유의적인 성장곡선 비율

변수	N	1차 방정식		2차 방정식	
		Valid n	n/N	Valid n	n/N
ALT	1030	188	18.25%	109	10.58%
ALB	1030	274	26.60%	124	12.03%
BMI	1030	188	18.25%	120	11.65%
CHOL	1030	131	12.71%	96	9.32%
CO2	1030	98	9.51%	135	13.10%
GLUCOSE	1030	153	14.85%	118	11.45%
HDL	1030	186	18.05%	117	11.35%
T4	1030	56	5.43%	66	6.40%
Trig	1030	118	11.45%	96	9.32%
Uric	1030	131	12.71%	117	11.35%
FVC	1030	140	13.59%	118	11.45%
Mean			14.67%		10.72%

$X_{iT} = \alpha_i + \beta_i \times T + \varepsilon_{iT}$ 와 같은 1차 선형모형에 의해 적합된 Risk Factor들은 현재 알고 있는 4개 시점에서 값 이외에도 5번째 검사 시점에서의 수치 그 이상의 검사시점에서의 수치도 예상할 수도 있다. 1차 선형모형을 통해 추정된 5번째 시점에서의 Risk Factor의 수치와 실제 검진자가 검사하여 얻은 값간의 차이를 살펴보기 위하여 오차를 추정하였다. 먼저 각 검사항목에 대한 성장곡선 추정치의 값을 구할 수 있다.

식6. 성장곡선의 형태들

$$\hat{X}_i = \alpha_i + \beta_{i1} \times T$$

$$\hat{X}_i = \alpha_i + \beta_{i1} \times T + \beta_{i2} \times T^2$$

$$\hat{X}_i = \alpha_i + \beta_{i1} \times \sqrt{T}$$

$$\hat{X}_i = \alpha_i + \beta_{i1} \times \log T$$

식7. 오차율

$$E_i = \frac{X_i - \hat{X}_i}{X_i}$$

단, X = 건강검진을 통해 실제 구한값,

$\hat{X}$ = 일차 선형모형을 통해 추정된 값

일차 선형 모형에 의한 성장 곡선을 통해 5번째 시점을 예측해 볼 수 있었으며 이 예측이 얼마나 정확한지에 대해 오차율을 구해서 생각해 볼 수 있다. 먼저 2차 선형모형의 경우 유의적인 성장곡선의 수가 비교적 적게 나타나고 있어 분석 대상에서 제외 시켰으며 1차 선형모형에 적합 시켜 5번째 시점의 수치를 예측한 값들에 대한 오차율이 표7과 나타나고 있다.

표7. 5번째 시점에서의 평균 오차율

간기능변수	N	P-value < 0.1	5th N	평균오차율	positive N	평균오차율	negative N	평균오차율
ALT	1030	188	82	0.466	26	0.434	56	0.481
ALB	1030	274	106	0.028	35	0.028	71	0.029
BMI	1030	188	80	0.595	18	0.183	59	0.720
CHOL	1030	131	44	0.065	10	0.088	34	0.058
CO2	1030	98	29	0.056	9	0.045	20	0.061
GLUCOSE	1030	153	65	0.060	17	0.109	48	0.042
HDL	1030	186	73	0.100	26	0.104	47	0.098
T4	1030	56	18	0.093	7	0.079	11	0.102
TRIG	1030	118	45	0.165	11	0.191	34	0.157
URIC	1030	131	60	0.079	16	0.074	44	0.081
FVC	1030	140	65	0.033	19	0.035	47	0.033

먼저 4개의 시점을 가지고 있는 1030명의 검진자중 유의적인 (P-value < 0.1) 기울기를 가지고 있는 성장곡선의 수는 Risk Factor에 따라 약간의 차이는 있지만 대략 56 ~ 274개가 존재하고 있음을 표를 통해 살펴볼 수 있다. 유의적인 성장곡선의 기울기를 가진 건강검진자 중 실제 5번째 건강검진을 받은 수는(5th N) 29 ~106명으로 이들을 통해 평균오차율을 구해 볼 수 있었다. Risk Factor에 따라 2.8% ~ 59.5%로 많은 차이를 나타내고 있다. 반면 간 질환이 있는 경우와 없는 경우 오차율의 차는 별로 나타나고 있지 않음을 표를 통해 살펴 볼 수 있었다. 더 나아가 6번째 건강검진을 받는다는 가정 하에서 실제 추정된 Risk Factor의 값과 실제 값과의 오차율을 아래의 표에서 살펴볼 수 있다.

표8. 6번째 시점에서의 평균 오차율

간기능변수 N		P-value < 0.1	6th N	평균오차율	positive N	평균오차율	negative N	평균오차율
ALT	1030	188	29	0.828	10	0.620	19	0.938
ALB	1030	274	36	0.054	15	0.058	21	0.051
BMI	1030	188	0	.	0	.	0	.
CHOL	1030	131	18	0.128	4	0.131	14	0.127
CO2	1030	98	6	0.091	1	0.065	5	0.096
GLUCOSE	1030	153	29	0.094	10	0.077	19	0.103
HDL	1030	186	27	0.188	8	0.173	19	0.194
T4	1030	56	5	0.177	2	0.084	3	0.240
TRIG	1030	118	22	0.293	6	0.273	16	0.301
URIC	1030	131	30	0.163	10	0.143	20	0.173
FVC	1030	140	0	.	0	.	0	.

추정된 6번째 시점의 오차율을 살펴보면 대략 5.4% ~ 82.8%로 Risk Factor에 따라서는 오차율이 상당히 적게 나타나고 있음을 확인해 볼 수 있다.

5.2.2 $\sqrt{X}$와 $\log X$에 대한 성장 곡선 모형

1차 및 2차 모형에 적합 시킨 성장곡선 분석을 통해 유의적인 곡선의 수와 오차율을 통해 사용가능 정도에 대해 살펴보았다. 다음으로는 $\sqrt{X}$와 $\log X$의 모형에 적합시킨 성장곡선 분석 결과를 살펴보겠다. 먼저 $\sqrt{X}$와 $\log X$의 유의적인 성장곡선의 비율은 평균 15.27%와 15.89%로 일차 이차 선형모형에 적합 시켰을 경우에 비해 유의적인 기울기를 가진 수가 많이 나타나고 있음을 볼 수 있다. (표 9) 간기능 수치인 ALB와 ALT와 같은 항목의 경우는 유의적인 기울기를 가진 성장곡선의 비율이 전체에 대해 20%이상을 나타내고 있음을 볼 수 있다.

표9. 유의적인 성장곡선 비율

변수	N	$\sqrt{X}$방정식		$\log X$ 방정식	
		Valid n	n/N	Valid n	n/N
ALT	1030	210	20.38%	226	21.91%
ALB	1030	275	26.69%	232	22.52%
BMI	1030	178	17.28%	170	16.50%
CHOL	1030	127	12.33%	123	11.94%
CO2	1030	104	10.09%	94	9.12%
GLUCOSE	1030	156	15.14%	155	15.04%
HDL	1030	189	18.34%	189	18.34%
T4	1030	91	8.83%	229	22.23%
Trig	1030	119	11.55%	112	10.87%
Uric	1030	141	13.68%	130	12.62%
FVC	1030	141	13.68%	142	13.78%
mean			15.27%		15.89%

이러한 유의적인 기울기를 가진 성장곡선에 대해 오차율을 살펴볼 수 있다. 성장곡선 분석을 통해 추정된 5번째 시점에서의 수치와 실제 수치간에 오차율의 경우 ALB와 같은 간 기능 수치는 오차율이 1.4%로 성장곡선을 통해 예측된 값과 실제 검진을 받은 값간에 거의 차이를 없음을 살펴볼 수 있으며 그 외에 Fvc와 CO2와 같은 검진항목에서도 낮은 오차율을 나타내고 있음을 볼 수 있다. 반면 BMI, T4, ALT와 같은 검진항목에서는 많은 오차율을 나타내고 있음을 살펴볼 수 있다.

표10. Log X 함수에서의 5번째 시점에서의 평균 오차율

간기능변수	N	P-value < 0.1	5th N	평균오차율	positive N	평균오차율	negative N	평균오차율
ALT	1030	226	86	0.223	26	0.234	60	0.218
ALB	1030	232	92	0.014	34	0.012	58	0.015
BMI	1030	170	76	0.378	21	0.215	56	0.436
CHOL	1030	123	40	0.030	12	0.037	28	0.027
CO2	1030	94	25	0.041	8	0.039	17	0.043
GLUCOSE	1030	155	67	0.034	22	0.051	45	0.026
HDL	1030	189	76	0.045	27	0.046	49	0.045
T4	1030	229	103	0.267	29	0.261	74	0.269
TRIG	1030	112	42	0.111	10	0.142	32	0.101
URIC	1030	130	55	0.047	17	0.045	38	0.048
FVC	1030	142	63	0.013	21	0.013	42	0.073

더 나아가 6번째 건강검진에서의 추정된 값과 실제 값의 오차율을 살펴보면 ALB와 같은 수치는 오차율이 2.4%로 낮게 나타나고 있다. 그러나 특이하게 5번째 시점에서 높은 오차율을 나타내던 BMI 수치가 낮은 오차율 (5.7%)를 나타내고 있는데 이는 그 분석대상의 수가 16개로 너무 작게 나타나고 있어서 다시 Sample Size를 늘린 후 다시 고려해 봐야할 것으로 생각되어진다.

표11. Log X 함수에서의 6번째 시점에서의 평균 오차율

간기능변수	N	P-value < 0.1	6th N	평균오차율	positive N	평균오차율	negative N	평균오차율
ALT	1030	226	31	0.350	11	0.326	20	0.363
ALB	1030	232	28	0.024	11	0.021	17	0.025
BMI	1030	170	16	0.057	3	0.063	13	0.056
CHOL	1030	123	0	.	0	.	0	.
CO2	1030	94	2	0.048	1	0.086	1	0.010
GLUCOSE	1030	155	30	0.046	13	0.054	17	0.039
HDL	1030	189	18	0.196	5	0.310	13	0.153
T4	1030	229	0	.	0	.	0	.
TRIG	1030	112	0	.	0	.	0	.
URIC	1030	130	24	0.083	7	0.084	17	0.083
FVC	1030	142	0	.	0	.	0	.

다음으로 $\sqrt{X}$함수에 적합 시킨 성장곡선에서의 오차율을 살펴볼 수 있다. 먼저 ALB와 같은 수치의 경우 오차율이 1.9%로 상당히 낮게 나타나고 있으며 Fvc 2.2% T4 2.0%등으로 다른 성장곡선에 비해 대체적으로 작은 오차율을 나타내고 있음을 아래의 표를 통해 살펴볼 수 있다.

표12. $\sqrt{X}$ 함수에서의 5번째 시점에서의 평균 오차율

간기능변수 N		P-value < 0.1	5th N	평균오차율	positive N	평균오차율	negative N	평균오차율
ALT	1030	210	87	0.331	29	0.319	58	0.337
ALB	1030	275	104	0.019	35	0.019	69	0.020
BMI	1030	178	79	0.419	19	0.130	60	0.511
CHOL	1030	127	42	0.045	11	0.055	31	0.042
CO2	1030	104	31	0.048	11	0.040	20	0.053
GLUCOSE	1030	156	68	0.041	20	0.064	48	0.032
HDL	1030	189	74	0.066	24	0.061	50	0.069
T4	1030	91	33	0.200	10	0.171	23	0.213
TRIG	1030	119	45	0.150	11	0.205	34	0.133
URIC	1030	141	64	0.066	18	0.067	46	0.066
FVC	1030	141	62	0.022	20	0.023	42	0.022

또한 6번째 시점에서 예측된 추정값과 실제 값과의 오차율을 살펴보면 ALB 수치의 경우 3.6%, Co2 6.4%로 낮은 오차율을 나타내고 있음을 볼 수 있다.

표13. $\sqrt{X}$ 함수에서의 6번째 시점에서의 평균 오차율

간기능변수	N	P-value < 0.1	6th N	평균오차율	positive N	평균오차율	negative N	평균오차율
ALT	1030	210	33	0.537	12	0.427	21	0.601
ALB	1030	275	34	0.036	13	0.039	21	0.034
BMI	1030	178	0	.	0	.	0	.
CHOL	1030	127	16	0.088	3	0.100	13	0.085
CO2	1030	104	5	0.064	1	0.119	4	0.050
GLUCOSE	1030	156	28	0.067	11	0.073	17	0.063
HDL	1030	189	25	0.128	7	0.116	18	0.133
T4	1030	91	8	0.347	3	0.205	5	0.432
TRIG	1030	119	20	0.267	6	0.404	14	0.208
URIC	1030	141	29	0.126	9	0.110	20	0.133
FVC	1030	141	0	.	0	.	0	.

제 6장 성장곡선을 이용한 여러 가지 판별 모형들

성장곡선 분석을 통해 추정된 값들을 여러 가지 간 질환 판별 예측모형에 사용하였을 경우 모형의 적합성 및 Risk Factor에 대해 살펴볼 수 있다. 먼저 4번 이상의 건강검진을 받은 1030명을 분석대상으로 미래에서의 (5번째 시점) 에서의 간 질환을 예측하는 모형을 추정해 보고자 한다. 성장곡선 분석을 통해 유의적인 기울기를 가진 Risk Factor에 대해서는 추정치를 사용하였으며 만약 성장곡선 분석에서 유의적이지 않은 경우에 대해서는 4번째 시점에서 검진 받은 자료를 기초로 분석을 하였다. 이렇게 1030명의 자료(Training Set)를 통해 간 질환 분석모형을 로지스틱 회귀모형, 나무모형, 신경망 모형을 통해 추정하였으며 1030명 중 실제 5번째 시점에서 검사를 받은 430명(Validation Set)을 통해 모형의 적합성에 대해 평가를 하였다.

6.1. Logistic Regression

추정할 수 있는 간 질환 예측 모형중 하나가 로지스틱 회귀모형이다.
일반적으로 로지스틱 회귀모형의 형태 아래와 같이 표현할 수 있다.

식8. Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times X_{ij} + \beta_{i2} \times X_{ij} + \beta_{i3} \times X_{ij} + \cdots + \varepsilon_i]$$

그러나 이러한 판별모형에서 사용되어지는 독립변수가 성장곡선을 통해 얻어

진 결과를 통해 사용되어질 때 이러한 판별모형의 형태는 변하게 되며 아래와 같이 표현되어진다. 즉 로지스틱 회귀모형에서 독립변수 X가 성장곡선을 통해 추정된 값 $\hat{X}$ ($= f(X)$)이 사용되어질 수 있으며 함수 f는 일차선형함수, 또는 log 함수, root 함수 등이 되어질 수 있다.

식9. 성장곡선 분석을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times f(X_{ij1}) + \beta_{i2} \times f(X_{ij2}) + \beta_{i3} \times f(X_{ij3}) + \cdots + \varepsilon_i]$$

단, $f(X_{mj}) = \alpha_m + \beta_{m1} \times T$

or $f(X_{mj}) = \alpha_m + \beta_{m1} \times \sqrt{T}$

or $f(X_{mj}) = \alpha_m + \beta_{m1} \times \log T$

or $f(X_{mj}) = X_{(4)j}$

식10. 일차성장곡선을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times (\alpha_m + \beta_{m11} \times T) + \beta_{i2} \times (\alpha_m + \beta_{m21} \times T) + \beta_{i3} \times (\alpha_m + \beta_{m31} \times T) + \cdots + \varepsilon_i]$$

식11. $\sqrt{X}$ 성장곡선을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times (\alpha_m + \beta_{m11} \times \sqrt{T}) + \beta_{i2} \times (\alpha_m + \beta_{m21} \times \sqrt{T}) + \beta_{i3} \times (\alpha_m + \beta_{m31} \times \sqrt{T}) + \cdots + \varepsilon_i]$$

식12. Log X 성장곡선을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times (\alpha_m + \beta_{m11} \times \log T) + \beta_{i2} \times (\alpha_m + \beta_{m21} \times \log T) + \beta_{i3} \times (\alpha_m + \beta_{m31} \times \log T) + \cdots + \varepsilon_i]$$

식13. 4번째 측정값을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times X_{(4)1j} + \beta_{i2} \times X_{(4)2j} + \beta_{i3} \times X_{(4)3j} + \cdots + \varepsilon_i]$$

위와 같이 추정된 로지스틱 모형을 통해 간 질환 판별 여부에 대해 살펴볼 수 있는데 주로 관심을 가질 수 있는 부분은 이미 규정되어진 식에 의한 회귀계수들과 모형의 적합성에 대한 것이다. 회귀계수를 통해 간 질환에 영향을 주는 항목의 정도를 살펴볼 수 있으며 모형의 적합성 부분에서는 accuracy와 sensitivity, specificity의 적당한 수준을 추정할 수 있다. 성장곡선 분석을 통해 얻은 값이 독립변수로 사용되어진 각각의 로지스틱 회귀분석에 대해 성장곡선 분석 없이 사용되어진 로지스틱 분석과의 차이점에 대해 살펴보겠다.

6.1.1 4번째 시점을 통해 추정된 값의 Logistic 간 질환 예측 모형

어떤 검진자가 4번의 검진을 받고 앞으로 5번째 검진을 받았을 때 간 질환에 걸릴지에 대해 관심을 가질 수 있다. 이러한 예측은 4번째 시점에서 간 질환에 영향을 주는 Risk Factor에 대한 분석으로 추정할 수 있으며 이를 아래와 같은 로지스틱 회귀모형의 표현식으로 나타낼 수 있다.

식14. 4번째 측정값을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times f(X_{1ij}) + \beta_{i2} \times f(X_{2ij}) + \beta_{i3} \times f(X_{3ij}) + \cdots + \varepsilon_i]$$

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times X_{(4)1j} + \beta_{i2} \times X_{(4)2j} + \beta_{i3} \times X_{(4)3j} + \cdots + \varepsilon_i]$$

표14. 4번째 측정값을 통한 Logistic Regression 결과

Parmeter	Estimate	SE	Wald chi-square	95% C.I. Lower	95% C.I. upper	p-value
Intercept	-3.1463	0.8316	14.31	-4.7761	-1.5164	0.0002
Fvc	-0.0146	0.0058	6.26	-0.0260	-0.0031	0.0124
BMI	0.0661	0.0077	72.64	0.0509	0.0813	0.0001
BioCreat	-1.0920	0.5069	4.64	-2.0856	-0.0984	0.0312
Biouric	0.2599	0.0726	12.82	0.1176	0.4022	0.0003
BioAlt	0.0211	0.0051	16.60	0.0109	0.0312	0.0001
BioLdh	-0.0016	0.0007	4.44	-0.0032	-0.0001	0.0351
BioTrig	0.0053	0.0009	28.75	0.0033	0.0072	0.0001
age	0.0379	0.0088	18.31	0.0206	0.0553	0.0001

분석결과 신체계측지수(BMI)와 폐기능 수치인 노력성 폐활량(Fvc) 생화학 검사시 대사 및 전해질 중 하나인 크레아티닌 수치(Creat), 간기능 수치인 ALT, 요산 수치(bioUric), 혈청지질 수치(BioTrig, BioHdl) 등이 간질환에 영향을 주는 주

요한 Risk Factor로 확인되었다. 이러한 Risk Factor 들은 이미 발표된 간암 관련 논문들에서 언급된 항목들이며 간 질환 발생에도 주요한 영향을 주는 항목들로 간주되어진다. 이러한 Risk Factor로 추정되어진 로지스틱 회귀모형을 살펴보면 cut-off value 0.4 수준에서 간질환 모형의 accuracy가 70.69%로 비교적 높게 나타나고 있으며 간 질환을 모형에 의해 맞추는 확률, 즉 Sensitivity가 74.48%로 비교적 높게 나타나고 있다.

표15 4번째 측정값을 통한 Logistic Regression Classification Table

	Predicted Probability	specificity	sencitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	24.70%	96.17%	75.30%	3.83%	50.09%	49.01%
	0.2	49.85%	92.08%	50.15%	7.92%	64.85%	35.15%
	0.3	65.51%	83.06%	34.49%	16.94%	71.74%	28.26%
	0.4	77.86%	67.49%	22.14%	32.51%	74.17%	25.83%
	0.5	85.99%	52.73%	14.01%	47.27%	74.17%	25.83%
	0.6	92.32%	42.35%	7.68%	57.65%	74.56%	25.44%
	0.7	95.63%	30.33%	4.37%	69.67%	72.42%	27.58%
	0.8	98.49%	17.21%	1.51%	82.79%	69.61%	30.39%
	0.9	99.25%	6.56%	0.75%	93.44%	66.31%	33.69%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	28.42%	93.10%	71.58%	6.90%	50.23%	49.77%
	0.2	51.93%	86.21%	48.07%	13.79%	64.38%	35.62%
	0.3	68.77%	74.48%	31.23%	25.52%	70.69%	29.31%
	0.4	78.60%	61.38%	21.40%	38.62%	72.79%	27.21%
	0.5	87.37%	46.90%	12.63%	53.10%	73.72%	26.28%
	0.6	94.39%	35.17%	5.61%	64.83%	74.41%	25.59%
	0.7	96.49%	24.14%	3.51%	75.86%	72.09%	27.91%
	0.8	98.25%	14.48%	1.75%	85.52%	70.00%	30.00%
	0.9	99.30%	4.83%	0.70%	95.17%	67.44%	32.56%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.1.2 1차 성장곡선을 통한 Logistic 간 질환 예측 모형

성장곡선을 통해 구한 앞으로의 유의적인 변화를 가지는 Risk Factor의 추정치들을 통해 로지스틱에 의한 간 질환 예측모형을 아래와 같이 추정해 볼 수 있다. 먼저 성장곡선을 통해 간 질환에 영향을 줄 수 있는 건강검진 항목을 일차 선형모형에 적합시키면 $f(X_{mj}) = \alpha_m + \beta_{m1} \times T + \beta_{m2}$ 같은 건강검진 항목의 추정치들을 구할 수 있다. 이 값을 로지스틱 모형에 대입하여 아래와 같은 일차선형모형 성장곡선을 통한 간 질환 예측모형을 추정해 볼 수 있다.

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times f(X_{1ij}) + \beta_{i2} \times f(X_{2ij}) + \beta_{i3} \times f(X_{3ij}) + \cdots + \varepsilon_i]$$

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times (\alpha_m + \beta_{m11} \times T) + \beta_{i2} \times (\alpha_m + \beta_{m21} \times T) + \beta_{i3} \times (\alpha_m + \beta_{m31} \times T) + \cdots + \varepsilon_i]$$

표16 1차 성장곡선에 의한 Logistic Regression 결과

Parmeter	Estimate	SE	Wald chi-square	95% C.I. Lower	95% C.I. upper	p-value
Intercept	-3.9816	0.6880	33.49	-5.3301	-2.6332	0.0001
Age	0.0366	0.0087	17.56	0.0195	0.0537	0.0001
BioAlt	0.0197	0.0045	18.61	0.0108	0.0287	0.0001
BMI	0.0558	0.0073	57.85	0.0414	0.0702	0.0001
BioHdl	-0.0179	0.0064	7.63	-0.0306	-0.0052	0.0057
Biotrig	0.0036	0.0008	17.64	0.0019	0.0054	0.0001
Biouric	0.1513	0.0586	6.67	0.0365	0.2662	0.0098

성장곡선 분석시 일차 선형모형에 적합시킨 로지스틱 회귀분석에서 유의적인 Risk Factor를 보면 연령과 간기능 수치인 ALT, 신체계측지수, 혈청지질 수치들이 유의적인 것으로 나타나고 있다. BioUric과 같은 검사항목이 간 질환 발생에 가장 많은 영향을 주는 요인으로 나타나고 있음을 살펴볼 수 있으며 이러한 모형에 의해 cut-off value 0.3 수준에서 전체 자료의 약 68.83%를 예측할 수 있으며

간 질환자에 대한 예측력은 81.38%로 상당히 높게 나타나고 있음을 살펴볼 수 있다. 또한 한가지 흥미로운 점은 모형 추정을 위한 Training Set에서의 accuracy (68.54%)보다는 Validation Set에서의 accuracy(68.83%)가 높게 나타나고 있음을 살펴볼 수 있다.

표17. 1차 성장곡선을 통한 Logistic Regression Classification Table

	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	20.33%	96.99%	79.67%	3.01%	47.57%	52.43%
	0.2	43.98%	90.71%	56.02%	9.29%	60.58%	39.42%
	0.3	61.45%	81.42%	38.55%	18.58%	68.54%	31.46%
	0.4	75.30%	65.30%	24.70%	65.30%	71.74%	28.26%
	0.5	85.99%	50.55%	14.01%	49.45%	73.39%	26.61%
	0.6	93.52%	36.07%	6.48%	63.93%	73.10%	26.90%
	0.7	97.14%	23.77%	2.86%	76.23%	71.06%	28.94%
	0.8	99.10%	12.57%	0.90%	87.43%	68.34%	31.66%
	0.9	99.40%	3.83%	0.60%	96.17%	65.43%	34.57%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	24.91%	95.17%	75.09%	4.83%	48.60%	51.40%
	0.2	43.51%	91.03%	56.49%	8.97%	59.53%	40.47%
	0.3	62.46%	81.38%	37.54%	18.62%	68.83%	31.17%
	0.4	76.49%	62.76%	23.51%	37.24%	71.86%	28.14%
	0.5	85.61%	46.90%	14.39%	53.10%	72.55%	27.45%
	0.6	92.63%	31.72%	7.37%	68.28%	72.09%	27.91%
	0.7	96.49%	19.31%	3.51%	80.69%	70.46%	29.54%
	0.8	98.60%	12.41%	1.40%	87.59%	69.53%	30.47%
	0.9	99.30%	4.14%	0.70%	95.86%	67.20%	32.80%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.1.3 Log X 성장곡선 분석을 통한 Logistic 간 질환 예측모형

Log X 모형에 적합시킨 성장곡선과 같은 경우 유의적인 기울기를 가진 곡선의 수가 비교적 많이 나타나고 있었으며 또한 실제 값과의 오차율도 낮게 나타나고 있었다. 성장곡선에 의해 추정된 값을 로지스틱 모형에 대입할 경우 아래와 같은 산식으로 표현되어진다.

식15. Log X 성장곡선을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times f(X_{1ij}) + \beta_{i2} \times f(X_{2ij}) + \beta_{i3} \times f(X_{3ij}) + \cdots + \varepsilon_i]$$

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times (\alpha_m + \beta_{m11} \times \log T) + \beta_{i2} \times (\alpha_m + \beta_{m21} \times \log T) + \beta_{i3} \times (\alpha_m + \beta_{m31} \times \log T) + \cdots + \varepsilon_i]$$

먼저 로지스틱 모형에서 Age나 간기능 수치인 ALT, 신체계측 지수인 BMI, 요산(BioUric), 혈청지질 수치인 BioTrig가 유의적인 Risk Factor로 나타나고 있음을 아래의 결과를 통해 살펴볼 수 있다. 이중 요산(BioUric)이 간 질환 발생에 가장 많은 영향을 주고 있음을 살펴 볼 수 있다.

표18. Log X성장곡선을 통한 Logistic Regression 결과

Parmeter	Estimate	SE	Wald chi-square	95% C.I. Lower	95% C.I. upper	p-value
Intercept	-5.2594	0.6086	74.69	-6.4522	-4.0666	0.0001
Age	0.0363	0.0087	17.24	0.0192	0.0535	0.0001
BioAlt	0.0241	0.0051	21.89	0.0140	0.0341	0.0001
BMI	0.0608	0.0074	66.66	0.0462	0.0754	0.0001
Biotrig	0.0049	0.0009	26.93	0.0030	0.0067	0.0001
BioUric	0.1731	0.0603	8.25	0.0550	0.2912	0.0041

LOG 모형에 의해 추정된 값으로 추정된 로지스틱 회귀모형에 의한 간 질환 예측 모형의 경우 cut-off value 0.3 수준에서 68.37%의 예측력을 나타내고 있으며 간 질환을 예측할 확률 역시 82.07%로 비교적 높게 나타나고 있음을 살펴볼 수 있다.

표19. Log X 성장곡선을 통한 Logistic Regression Classification Table

	Predicted Probability	specificity	sensitiviy	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	21.69%	97.54%	78.31%	2.46%	48.64%	51.36%
	0.2	46.08%	89.62%	53.92%	10.38%	61.55%	38.45%
	0.3	62.59%	79.51%	37.05%	20.49%	68.83%	31.17%
	0.4	75.60%	66.12%	24.04%	33.88%	72.23%	27.77%
	0.5	87.50%	50.82%	12.50%	49.18%	74.46%	25.54%
	0.6	93.22%	36.89%	6.78%	63.11%	73.20%	26.80%
	0.7	97.29%	23.22%	2.71%	76.78%	70.97%	29.03%
	0.8	98.95%	13.93%	1.05%	86.07%	68.73%	31.27%
	0.9	99.25%	5.19%	0.75%	94.81%	65.82%	34.18%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	22.46%	95.17%	77.54%	4.83%	46.97%	53.03%
	0.2	41.75%	90.34%	58.25%	9.66%	58.13%	41.87%
	0.3	61.40%	82.07%	38.60%	17.93%	68.37%	31.63%
	0.4	74.74%	65.52%	25.26%	34.48%	71.62%	28.38%
	0.5	82.46%	49.66%	17.54%	50.34%	71.39%	28.61%
	0.6	92.28%	37.24%	7.72%	62.76%	73.72%	26.28%
	0.7	96.49%	22.76%	3.51%	77.24%	71.62%	26.28%
	0.8	97.89%	15.86%	2.11%	84.14%	70.23%	29.77%
	0.9	99.30%	5.52%	0.70%	94.48%	67.67%	32.33%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	32.33%

6.1.4 $\sqrt{X}$ 성장곡선을 통한 Logistic 간 질환 예측 모형

성장곡선분석을 통해 $\sqrt{X}$ 모형의 오차율 및 유의적인 곡선의 수가 다른 모형에 비해 많이 나타나고 있음을 살펴보았다. 성장곡선을 통해 추정된 로지스틱에 의한 간 질환 예측모형은 아래와 같이 표현되어진다.

식16. $\sqrt{X}$ 성장곡선을 통한 Logistic Regression

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times f(X_{1ij}) + \beta_{i2} \times f(X_{2ij}) + \beta_{i3} \times f(X_{3ij}) + \cdots + \varepsilon_i]$$

$$\theta_i = \exp[\alpha_i + \beta_{i1} \times (\alpha_m + \beta_{m11} \times \sqrt{T}) + \beta_{i2} \times (\alpha_m + \beta_{m21} \times \sqrt{T}) + \beta_{i3} \times (\alpha_m + \beta_{m31} \times \sqrt{T}) + \cdots + \varepsilon_i]$$

로지스틱 회귀모형의 분석결과를 살펴보면 먼저 Log X 모형에서 유의적이었던 Risk Factor 대부분이 $\sqrt{X}$ 모형에서도 나타나고 있으며 혈청지질 수치인 BioHdl이 추가되었으며 요산 BioUric가 제외되어 있다.

표20. $\sqrt{X}$ 성장곡선을 통한 Logistic Regression 결과

Parmeter	Estimate	SE	Wald chi-square	95% C.I. Lower	95% C.I. upper	p-value
Intercept	-3.2762	0.5944	30.38	-4.4412	-2.1111	0.0001
age	0.0367	0.0087	17.78	0.0197	0.0538	0.0001
BioAlt	0.0242	0.0048	24.80	0.0147	0.0337	0.0001
BMI	0.0602	0.0074	65.30	0.0456	0.0748	0.0001
BioHdl	-0.0197	0.0065	8.96	-0.0326	-0.0068	0.0028
BioTrig	0.0040	0.0009	18.64	0.0022	0.0058	0.0001

$\sqrt{X}$ 모형 성장곡선 분석을 통해 추정된 값을 사용한 간 질환 예측모형의 경우 cut-off value 0.3 수준에서 accuracy가 68.13%이며 간 질환을 예측하는 확률인 Sensitivity가 79.31%로 4번째 시점을 사용한 로지스틱 모형보다는 높은 예측력을 나타내고 있지만 1차선형 모형이나 Log 모형에 비해 떨어지고 있음을 살펴볼 수 있었다.

표21. $\sqrt{X}$ 성장곡선을 통한 Logistic Regression Classification Table

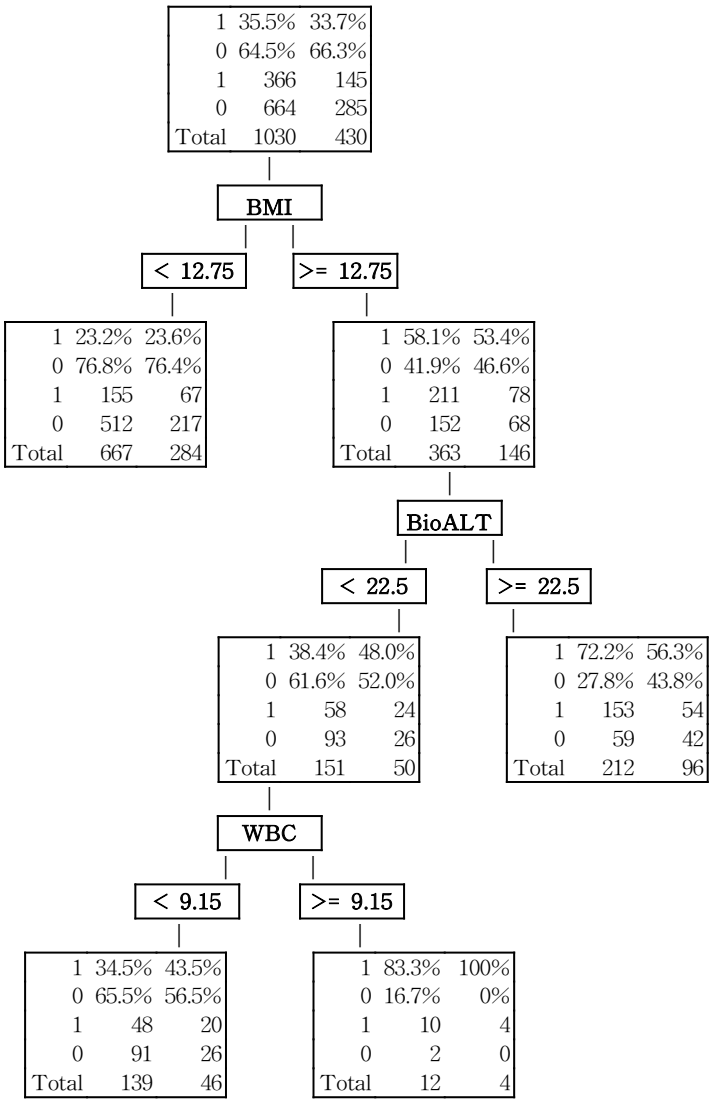
	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	19.43%	96.72%	80.57%	3.28%	46.89%	53.11%
	0.2	44.43%	90.16%	55.57%	9.84%	60.67%	39.33%
	0.3	61.90%	81.69%	38.10%	18.31%	68.93%	31.07%
	0.4	74.85%	65.85%	25.15%	34.15%	71.65%	28.35%
	0.5	85.84%	51.09%	14.16%	48.91%	73.49%	26.51%
	0.6	92.92%	36.34%	7.08%	63.66%	72.81%	27.19%
	0.7	97.59%	24.04%	2.41%	75.96%	71.45%	28.55%
	0.8	99.25%	12.02%	0.75%	87.98%	68.25%	31.75%
	0.9	99.40%	4.64%	0.60%	95.36%	65.72%	34.28%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	23.16%	95.17%	76.84%	4.83%	47.44%	52.56%
	0.2	44.56%	91.03%	55.44%	8.97%	60.23%	39.77%
	0.3	62.46%	79.31%	37.54%	20.69%	68.13%	31.87%
	0.4	75.79%	62.07%	24.21%	37.93%	71.16%	28.84%
	0.5	83.86%	47.59%	16.14%	52.41%	71.62%	28.38%
	0.6	92.98%	33.10%	7.02%	66.90%	72.79%	27.21%
	0.7	96.84%	21.38%	3.16%	78.62%	71.39%	28.61%
	0.8	98.60%	15.17%	1.40%	84.83%	70.46%	29.54%
	0.9	99.30%	4.14%	0.70%	95.86%	67.20%	32.80%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.2 성장곡선을 이용한 나무모형

6.2.1 4번째 시점시 나무모형을 통한 간 질환 예측모형

$\hat{X}_{(5)} = X_{(4)}$

그림6. 4번째 시점에 의한 나무모형



4번째 검사항목에 의해 추정된 나무모형과 같은 경우 BMI, ALT, WBC와 같은 항목에 의해 쉽고 간단한 예측 모형을 추정해 볼 수 있었다. 첫 번째 분리 기준인 BMI 수치가 12.75보다 작으면 간 질환에 대해 건강하다고 말 할 수 있으며 BMI 수치가 12.75보다 크고 ALT 수치 역시 22.5 보다 크면 간 질환에 걸릴 수 있다고 추정되어진다. 그 외에 BMI 수치가 12.75보다 크고 ALT 수치가 22.5 보다 작지만 WBC 수치가 9.15보다 크면 간 질환에 걸릴 가능성이 높다는 것을 나무 모형을 통해 살펴볼 수 있다. 이러한 나무 모형의 경우 accuracy와 간 질환 예측 확률인 Sensitivity가 너무 낮게 나타나고 있음을 확인할 수 있었다.

표22. 4번째 시점을 이용한 나무모형의 Classification Table

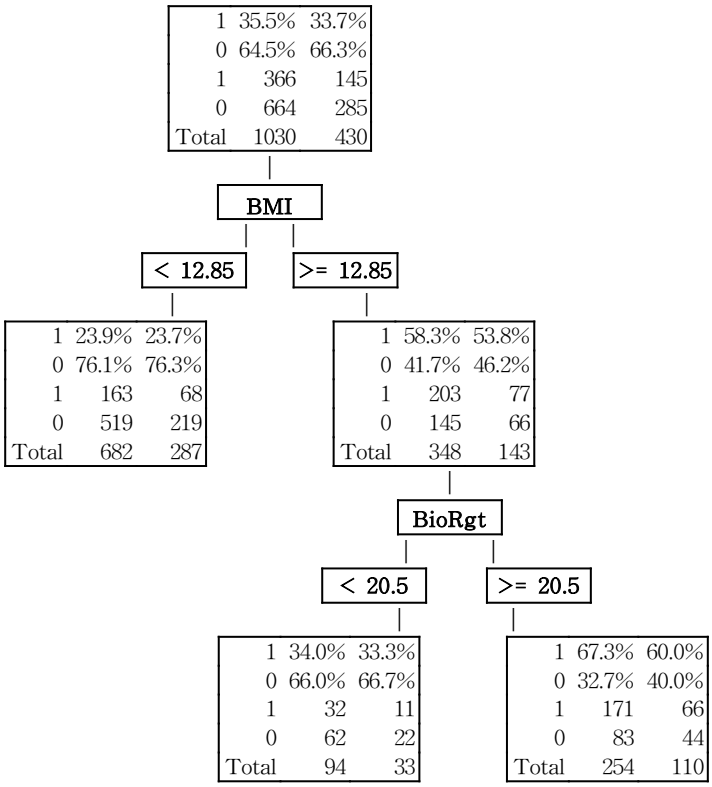
	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	0%	100.00%	100.00%	0.00%	35.53%	64.47%
	0.2	0%	100.00%	100.00%	0.00%	35.53%	64.47%
	0.3	77.11%	57.65%	22.89%	42.35%	70.19%	29.81%
	0.4	90.81%	44.54%	9.19%	55.46%	74.36%	25.64%
	0.5	90.81%	44.54%	9.19%	55.46%	74.36%	25.64%
	0.6	90.81%	44.54%	9.19%	55.46%	74.36%	25.64%
	0.7	90.81%	44.54%	9.19%	55.46%	74.36%	25.64%
	0.8	99.70%	2.73%	0.30%	97.27%	65.24%	34.76%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	0.00%	100.00%	100.00%	0.00%	66.27%	33.73%
	0.2	0.00%	100.00%	100.00%	0.00%	66.27%	33.73%
	0.3	76.14%	53.79%	23.86%	46.21%	68.60%	31.40%
	0.4	85.26%	40.00%	14.74%	60.00%	70.00%	30.00%
	0.5	85.26%	40.00%	14.74%	60.00%	70.00%	30.00%
	0.6	85.26%	40.00%	14.74%	60.00%	70.00%	30.00%
	0.7	85.26%	40.00%	14.74%	60.00%	70.00%	30.00%
	0.8	100.00%	2.76%	0.00%	97.24%	67.20%	32.80%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.2.2 1차 성장곡선 분석시 나무모형을 통한 간 질환 예측모형

$\hat{X}_{(5)} = \alpha_m + \beta_{m1} \times T + \beta_{m2}$ 만약 β_{i1}이 유의적일 경우

그 외 $\hat{X}_{(5)} = X_{(4)}$

그림7. 1차 성장곡선을 통한 나무모형



일차선형 성장곡선을 통해 추정된 값을 사용한 간 질환 예측모형을 살펴보면 먼저 BMI 수치가 12.85보다 작으면 간 질환에 대해 건강하다고 말할 수 있으며 만약 BMI 수치가 12.85보다 크지만 BioRgt 수치가 20.5 보다 작으면 비교적 간

질환에 건강하다고 간주되어진다. 반면 BMI 수치가 12.58보다 크며 BioRgt 역시 20.5보다 클 경우 간 질환에 걸릴 확률이 높은 위험군이라는 것을 나무모형을 통해 볼 수 있다. 이러한 나무 모형에 의한 간 질환 예측 모형의 accuracy와 간 질환자에 대한 예측 확률정도는 로지스틱 모형에 비해 많이 떨어져 나타나고 있음을 알 수 있다. 반면에 나무모형과 같은 경우 해석이 쉽고 용이하며 간 질환에 영향을 주는 Risk Factor의 기준값을 제시함으로써 결과에 대해 알기 쉽게 나타나고 있음을 살펴볼 수 있었다.

표23. 1차성장 곡선을 통한 나무모형의 Classification Table

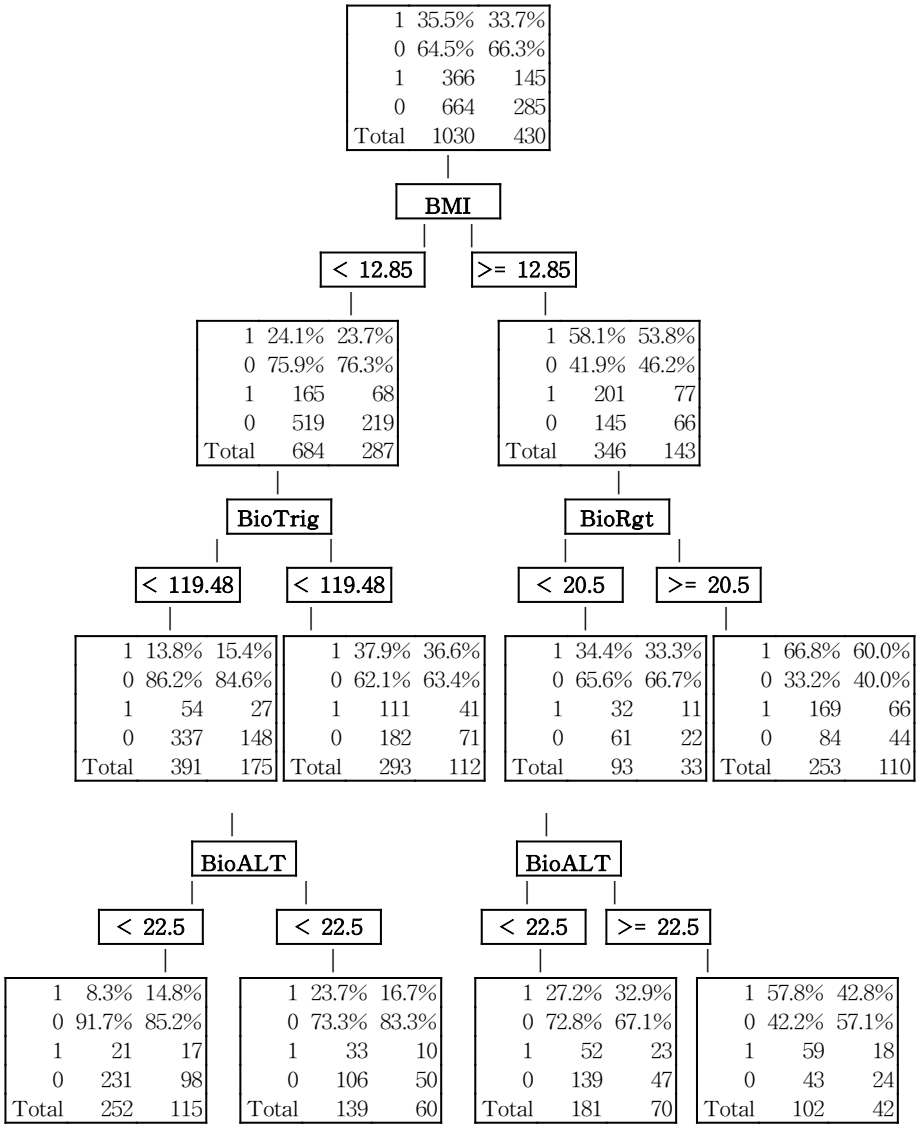
	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	0.00%	100.00%	100.00%	0.00%	35.53%	64.47%
	0.2	0.00%	100.00%	100.00%	0.00%	35.53%	64.47%
	0.3	78.16%	55.46%	21.84%	44.54%	70.09%	29.91%
	0.4	87.50%	46.72%	12.50%	53.28%	73.00%	27.00%
	0.5	87.50%	46.72%	12.50%	53.28%	73.00%	27.00%
	0.6	87.50%	46.72%	12.50%	53.28%	73.00%	27.00%
	0.7	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.8	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	0.00%	100.00%	100.00%	0.00%	33.72%	66.28%
	0.2	0.00%	100.00%	100.00%	0.00%	33.72%	66.28%
	0.3	76.84%	53.10%	23.16%	46.90%	68.83%	31.17%
	0.4	84.56%	45.52%	15.44%	54.48%	71.39%	28.61%
	0.5	84.56%	45.52%	15.44%	54.48%	71.39%	28.61%
	0.6	84.56%	45.52%	15.44%	54.48%	71.39%	28.61%
	0.7	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.8	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.2.3 Log X 성장곡선 분석시 나무모형을 통한 간 질환 예측모형

$\hat{X}_{(5)} = \alpha_i + \beta_{i1} \times \log T$ 만약 β_{i1}이 유의적일 경우

그 외 $\hat{X}_{(5)} = X_{(4)}$

그림8. Log X성장곡선을 통한 나무모형



Log X 성장곡선에 의한 나무 모형의 경우 그림에서 보여지듯이 BMI 수치와 BioTrig, BioRgt, BioAlt에 따라 간 질환 여부를 판단할 수 있었다. BMI 수치가 12.58보다 작고 BioTrig 역시 119.48 보다 작은 경우 간 질환에 대해 비교적 건강할 확률이 높게 나타나고 있으며 BMI 수치가 12.58보다 크고 BioRgt 수치가 20.5보다 클 경우 간 질환에 걸릴 확률이 높게 나타나고 있음을 확인해 볼 수 있었다. Log X 성장곡선에 의한 나무 모형의 경우 cut-off value 0.3 수준에서 validation Set의 accuracy가 69.53%로 비교적 높게 나타나고 있으며 Sensitivity 또한 67.59%로 다른 나무 모형들에 비해 비교적 우수하게 나타나고 있음을 확인해 볼 수 있었다.

표24. Log X 성장곡선을 통한 나무모형 Classification Table

	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	34.79%	94.26%	65.21%	5.74%	54.17%	45.83%
	0.2	67.47%	80.60%	32.53%	19.40%	72.13%	27.87%
	0.3	70.78%	78.42%	29.22%	21.58%	73.49%	26.51%
	0.4	79.97%	69.67%	20.03%	30.33%	76.31%	23.69%
	0.5	82.98%	64.21%	17.02%	35.79%	76.31%	23.69%
	0.6	84.04%	61.48%	15.96%	38.52%	76.01%	23.69%
	0.7	96.69%	15.30%	3.31%	84.70%	67.76%	32.24%
	0.8	99.85%	1.37%	0.15%	98.63%	64.85%	35.15%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	34.39%	88.28%	65.61%	11.72%	52.55%	47.45%
	0.2	66.32%	70.34%	33.68%	29.66%	67.67%	32.33%
	0.3	70.53%	67.59%	29.47%	32.41%	69.53%	30.47%
	0.4	78.25%	60.00%	21.75%	40.00%	72.09%	27.91%
	0.5	79.30%	56.55%	20.70%	43.45%	71.62%	28.38%
	0.6	79.30%	55.17%	20.70%	44.83%	71.16%	28.84%
	0.7	94.74%	9.66%	5.26%	90.34%	66.04%	33.96%
	0.8	98.95%	0.00%	1.05%	100.00%	65.58%	34.42%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.2.4 $\sqrt{X}$ 성장곡선 분석시 나무모형을 통한 간 질환 예측모형

$\hat{X}_{(5)} = \alpha_i + \beta_{i1} \times \sqrt{T}$ 만약 β_{i1}이 유의적일 경우

그 외 $\hat{X}_{(5)} = X_{(4)}$

그림9. $\sqrt{X}$성장곡선을 통한 나무모형

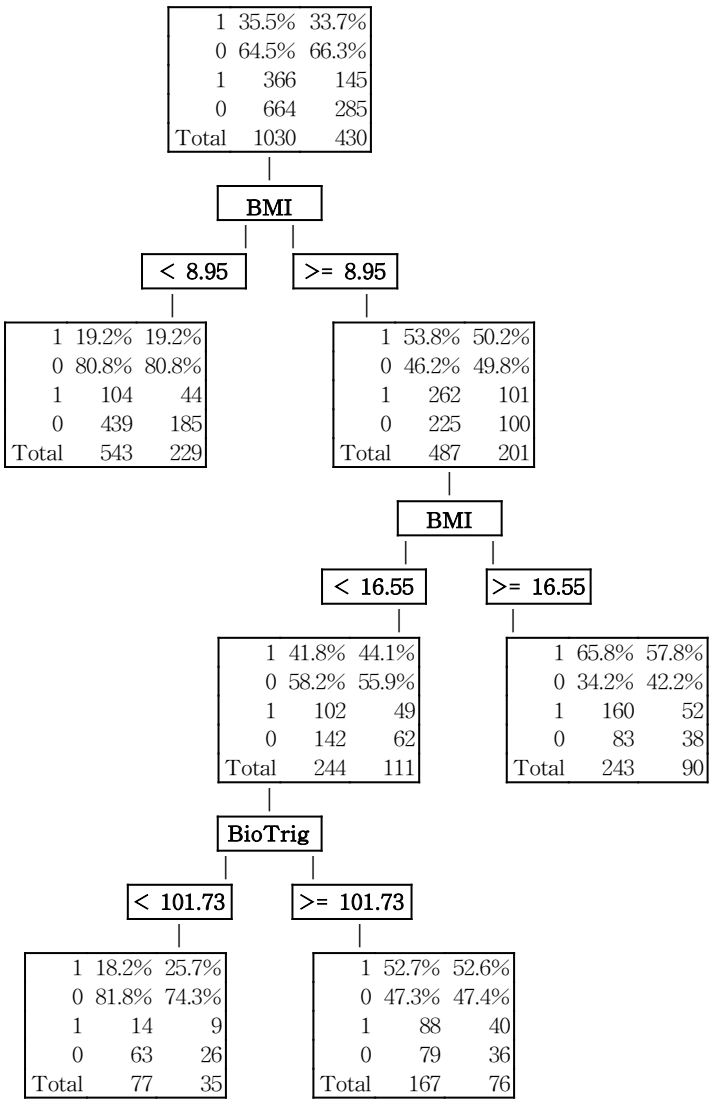


표25 $\sqrt{X}$ 성장곡선을 통한 나무모형 Classification Table

	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	0.00%	100.00%	100.00%	0.00%	35.53%	64.47%
	0.2	75.60%	67.76%	24.40%	32.24%	72.81%	27.19%
	0.3	80.27%	64.75%	19.73%	35.25%	74.75%	25.25%
	0.4	80.27%	64.75%	19.73%	35.25%	74.75%	25.25%
	0.5	80.27%	64.75%	19.73%	35.25%	74.75%	25.25%
	0.6	80.27%	64.75%	19.73%	35.25%	74.75%	25.25%
	0.7	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.8	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	0.00%	100.00%	100.00%	0.00%	33.72%	66.28%
	0.2	74.04%	63.45%	25.96%	36.55%	70.46%	29.54%
	0.3	76.14%	60.69%	23.89%	39.31%	70.93%	29.07%
	0.4	76.14%	60.69%	23.86%	39.31%	70.93%	29.07%
	0.5	76.14%	60.69%	23.86%	39.31%	70.93%	29.07%
	0.6	76.14%	60.69%	23.86%	39.31%	70.93%	29.07%
	0.7	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.8	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

$\sqrt{X}$에 의한 성장곡선 분석 결과를 이용하여 추정한 나무모형에서는 먼저 BMI 분리 기준이 8.95와 16.55 두 번 나타나게 되는데 이는 나무모형 추정시 binary split으로 분리기준을 정하여 분석하였기 때문이다. 먼저 BMI 수치가 8.95보다 적게 나타날 경우 간 질환에 대해 건강할 확률이 높게 나타나고 있으며 BMI 수치가 16.55보다 큰 경우는 간 질환에 걸릴 확률이 높게 나타나고 있다. 반면 BMI 수치가 8.95 보다 크고 16.55 보다는 작으면 BioTrig가 101.73 보다 작으면 간질환에 대해 비교적 건강하고 이 보다 크면 간 질환에 걸릴 확률이 비교적 높게 나타나고 있음을 확인할 수 있다. 이러한 나무모형에서 cut-off value 0.3 수준에서 다른 나무 모형에 비해 가장 높은 accuracy(70.93%)를 나타내고 있지만 Sensitivity(=60.69%)가 낮게 나타나고 있다.

6.3 성장곡선을 이용한 신경망 모형

Neural Network를 이용한 간 질환 예측 모형에서 건강검진 항목을 독립변수(입력층)하여 분석을 할 수 있다. 그러나 성장곡선을 통해 예측된 값을 독립변수의 추정치로 사용하여 Neural Network에 의한 간 질환 판별모형을 추정해 볼 수 있는데 먼저 일반적으로 신경망 모형의 함수를 다음과 같이 정의할 수 있다.

식17. 신경망 모형

$$y_k = g'\left(\sum_{j}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i\right)\right)$$

$g,\ g'$: activity function

w : weight

그러나 독립변수(입력층)에 대해 성장곡선을 이용한 추정치를 사용하여 Neural Network를 이용한 판별모형을 추정할 수 있으며 아래와 같은 산식의 전개과정을 통해 추정해 볼 수 있다.

식18. 성장곡선을 통한 신경망 모형

$$y_k = g'\left(\sum_{j}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)} \hat{x}_i\right)\right)$$

$$\hat{X}_i = X_{(4)}$$

$$\hat{X}_i = \alpha_i + \beta_{i1} \times T$$

$$\hat{X}_i = \alpha_i + \beta_{i1} \times \sqrt{T}$$

$$\hat{X}_i = \alpha_i + \beta_{i1} \times \log T$$

성장곡선을 통해 추정된 값을 통해 Neural Network에 의한 간 질환 판별 모형식은 다음과 같이 볼 수 있다. 첫 번째로 4번째 시점에서 건강검진을 받은 자료로써 5번째 시점에서의 간 질환 예측 모형의 추정치로 사용한 신경망 모형의 함수이다.

식19. 4번째 측정값에 의한 신경망 모형

$$y_k = g'\left(\sum_j^M w_{kj}^{(2)} g\left(\sum_{i=0}^d w_{ji}^{(1)} X_{(4)}\right)\right)$$

두 번째 모형은 일차선형모형에 의한 성장곡선 분석을 통해 추정된 값을 간 질환 예측 모형의 추정치로 사용한 신경망 모형의 함수이다.

식20. 일차 성장곡선에 의한 신경망 모형

$$y_k = g'\left(\sum_j^M w_{kj}^{(2)} g\left(\sum_{i=0}^d w_{ji}^{(1)}[\alpha_i + \beta_{i1} \times T]\right)\right)$$

세 번째 함수는 Log 모형에 의한 성장곡선 분석 추정치를 사용한 간 질환 예측 모형의 신경망 모형 함수이다.

식21. Log X성장곡선에 의한 신경망 모형

$$y_k = g'\left(\sum_j^M w_{kj}^{(2)} g\left(\sum_{i=0}^d w_{ji}^{(1)}[\alpha_i + \beta_{i1} \times \log T]\right)\right)$$

마지막으로 $\sqrt{X}$모형에 의한 성장곡선 분석 추정치를 사용한 간 질환 예측 모형의 신경망 모형 함수이다.

식22. $\sqrt{X}$ 성장곡선을 통한 신경망 모형

$$y_k = g'\left(\sum_{j}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)}[\alpha_i + \beta_{i1} \times \sqrt{T}]\right)\right)$$

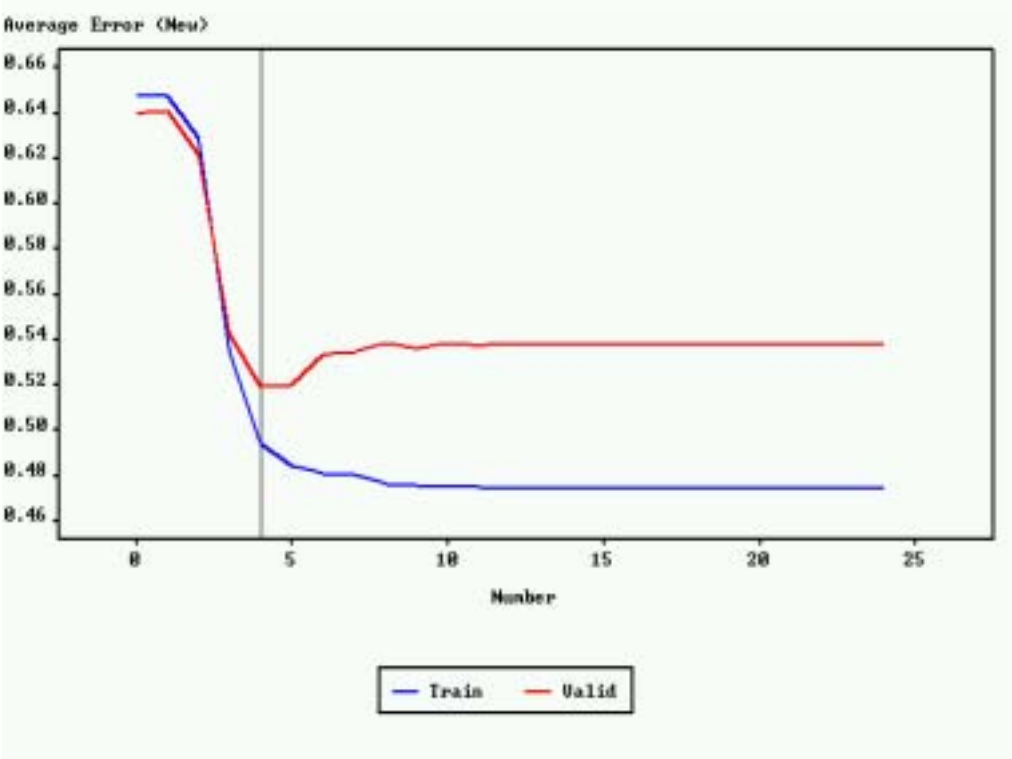
이렇게 추정된 성장곡선을 이용한 신경망 모형이 얼마나 간 질환을 정확히 예측하는지에 대해 Classification Table에 나타난 accuracy와 specificity를 통해 살펴 볼 수가 있다.

6.3.1 4번째 시점 기준시 신경망 분석을 통한 간 질환 예측모형

$$y_k = g'\left(\sum_{j}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)} X_{(4)}\right)\right)$$

4번째 시점에서 검진 받은 값을 통해 추정된 신경망 모형에 의한 간질환 예측모형에서 먼저 오차함수에 대한 반복회수를 나타낸 아래의 그림을 살펴보면 대략 iteration 4를 기준으로 validation set의 오차가 가장 최소를 나타내고 있음을 살펴볼 수 있다.

그림 10. 4번째 측정값에 의한 신경망 모형의 Average Error Plot



이와 같은 신경망 모형에 간 질환 예측모형의 accuracy를 살펴보면 cut-off value 0.3 수준에서 validation Set인 경우 72.09%를 나타내고 있으며 Sensitivity가 71.72%로 좋은 값을 나타내고 있음을 살펴볼 수 있다.

표26. 4번째 측정값에 의한 신경망 모형 Classification Table

	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	0.00%	100.00%	100.00%	0.00%	35.53%	64.47%
	0.2	60.39%	87.70%	39.61%	12.30%	70.09%	29.91%
	0.3	70.33%	80.33%	29.67%	19.67%	73.88%	26.12%
	0.4	78.77%	69.95%	21.23%	30.05%	75.63%	24.37%
	0.5	83.73%	62.02%	16.27%	37.98%	76.01%	23.99%
	0.6	88.55%	50.55%	11.45%	49.45%	72.13%	27.87%
	0.7	96.23%	25.41%	3.77%	74.59%	71.06%	28.94%
	0.8	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	0.00%	100.00%	100.00%	0.00%	33.72%	96.28%
	0.2	62.11%	78.62%	37.89%	21.38%	67.67%	32.33%
	0.3	72.28%	71.72%	27.72%	28.28%	72.09%	27.91%
	0.4	82.46%	60.69%	17.54%	39.31%	75.11%	24.89%
	0.5	85.26%	53.10%	14.74%	46.90%	74.41%	25.59%
	0.6	90.88%	42.76%	9.12%	57.24%	74.65%	25.35%
	0.7	96.84%	20.00%	3.16%	80.00%	70.93%	29.07%
	0.8	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

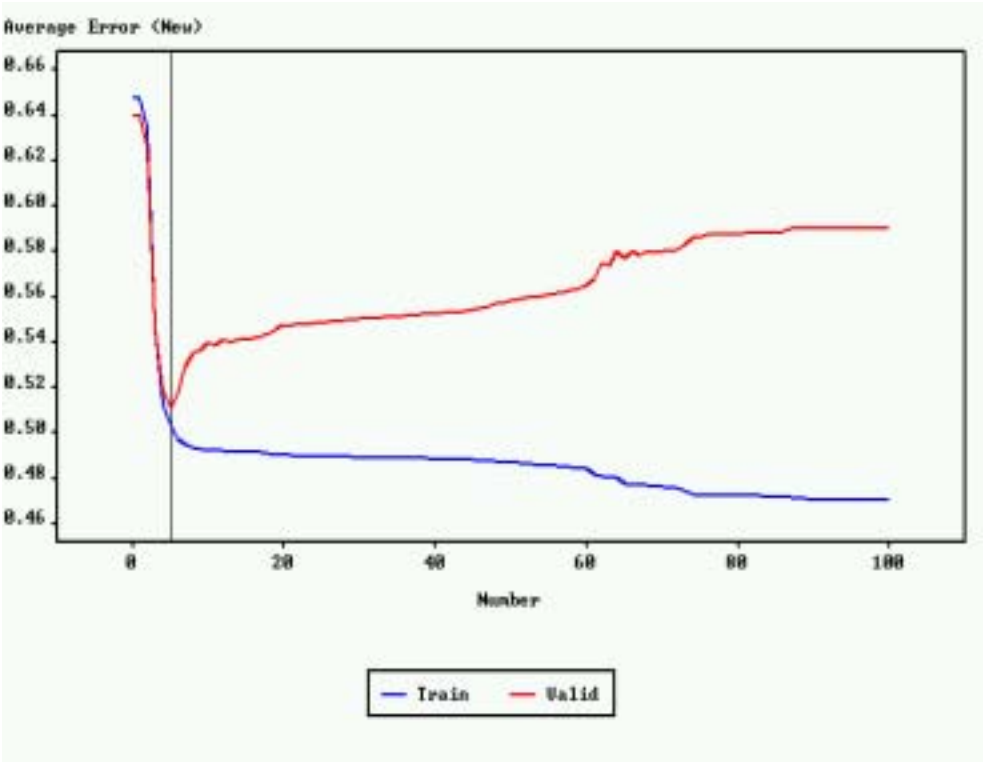
6.3.2 1차 성장곡선을 통한 신경망 분석을 통한 간 질환 예측모형

$$y_k = g'\left(\sum_j^M w_{kj}^{(2)} g\left(\sum_{i=0}^d w_{ji}^{(1)}[\alpha_i + \beta_{i1} \times T]\right)\right)$$

1차 선형 모형을 통한 성장곡선 분석을 통해 구한 미래시점의 Risk Factor의 추정치로 구한 신경망 모형에서는 iteration이 8인 경우 Validation Set의 오차가

최소를 나타내고 있음을 살펴볼 수 있다. iteration이 8 이상 증가되면 Validation Set의 오차가 점차적으로 증가하는 모습을 그래프를 통해 볼 수 있다.

그림 11. 일차 성장곡선에 의한 신경망 모형의 Average Error Plot



본 신경망 모형의 적합성을 살펴보면 Validation Set의 경우 cut-off value 0.3 수준에서 accuracy가 73.02를 나타내고 있으며 Sensitivity 역시 75.86으로 상당히 우수한 모형임을 알 수 있다.

표27. 1차성장곡선을 통한 신경망 모형의 Classification Table

	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	7.08%	99.45%	92.92%	0.55%	39.90%	60.10%
	0.2	57.83%	87.43%	42.17%	12.57%	68.34%	31.66%
	0.3	67.47%	80.87%	32.53%	19.13%	72.23%	27.77%
	0.4	76.20%	71.86%	23.80%	28.14%	74.66%	25.40%
	0.5	82.23%	62.30%	17.77%	37.70%	75.14%	24.86%
	0.6	88.10%	51.09%	11.90%	48.91%	74.95%	25.05%
	0.7	94.73%	28.42%	5.27%	71.58%	71.16%	28.84%
	0.8	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	7.72%	97.24%	92.28%	2.76%	37.90%	62.10%
	0.2	61.05%	84.83%	38.95%	15.17%	69.06%	37.94%
	0.3	71.58%	75.86%	28.42%	24.14%	73.02%	26.98%
	0.4	79.65%	62.76%	20.35%	37.24%	73.95%	26.05%
	0.5	82.81%	55.86%	17.19%	44.14%	73.72%	26.28%
	0.6	88.07%	45.52%	11.93%	54.48%	73.72%	26.28%
	0.7	95.79%	28.28%	4.21%	71.72%	73.02%	26.98%
	0.8	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

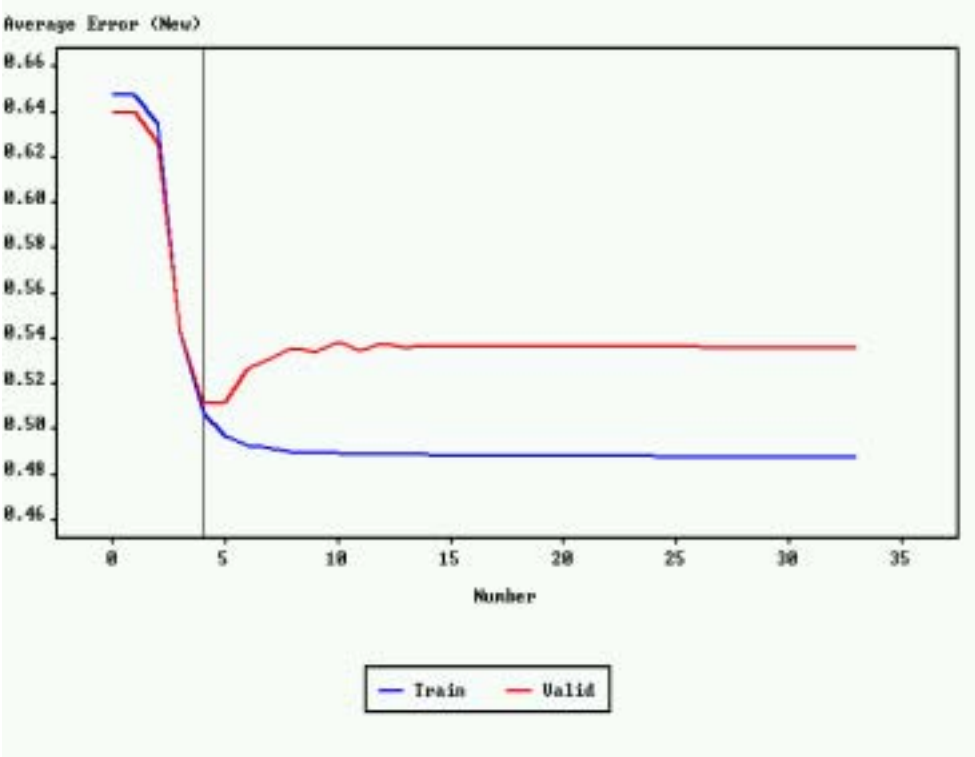
6.3.3 Log X 성장곡선 분석시 신경망 분석을 통한 간 질환 예측모형

$$y_k = g'\left(\sum_{j}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)}[\alpha_i + \beta_{i1} \times \log T]\right)\right)$$

Log 함수를 통한 성장곡선 분석 결과를 추정치로 사용한 신경망 모형에서는

iteration 4에서 Validation Set에서 오차가 최소를 나타내고 있음을 살펴볼 수 있다.

그림 12. Log X성장곡선에 의한 신경망 모형의 Average Error Plot



Log 성장곡선을 사용한 신경망 모형의 Validation Set의 accuracy를 살펴 보면 cut-off value 0.3 수준에서 72.79%로 높게 나타나고 있으며 Sensitivity 역시 76.55%로 상당히 높은 값을 나타내고 있음을 볼 수 있다.

표28. Log X성장곡선에 의한 신경망 모형의 Classification Table

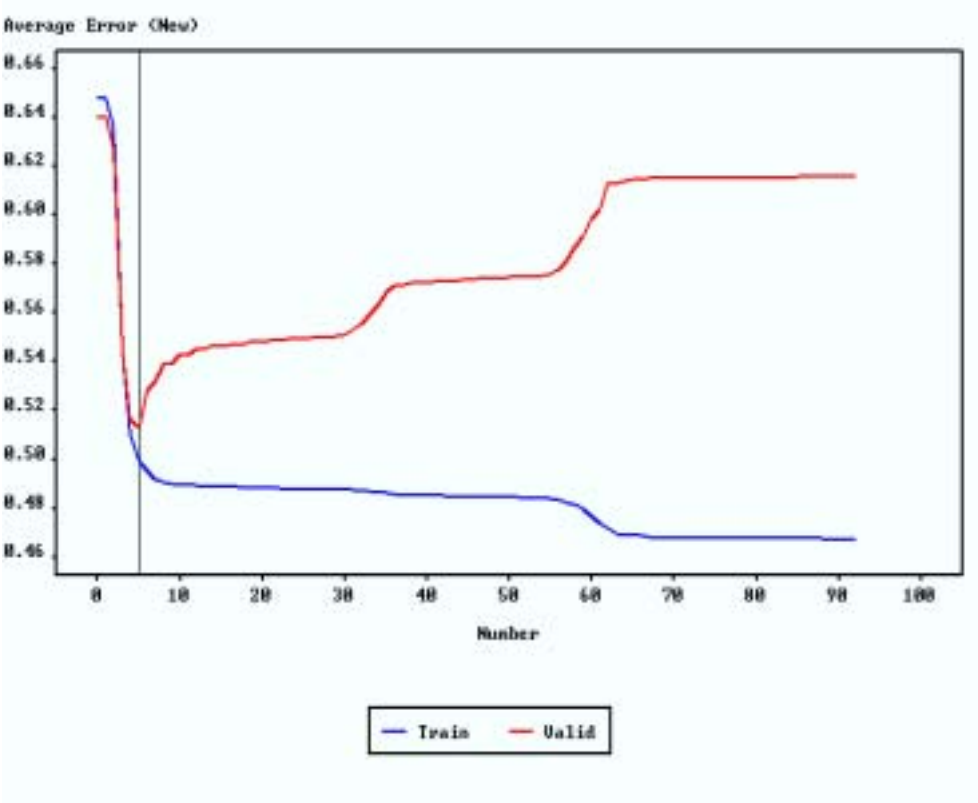
	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	0.00%	100.00%	100.00%	0.00%	35.53%	64.47%
	0.2	60.24%	85.79%	39.76%	14.21%	69.32%	30.68%
	0.3	69.13%	78.42%	30.87%	21.58%	72.42%	27.58%
	0.4	78.77%	68.03%	21.23%	31.97%	74.95%	25.05%
	0.5	84.04%	57.38%	15.96%	42.62%	74.56%	25.44%
	0.6	89.61%	44.54%	10.39%	55.46%	73.59%	26.41%
	0.7	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.8	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	0.00%	100.00%	100.00%	0.00%	33.72%	66.28%
	0.2	57.89%	84.83%	42.11%	15.17%	66.97%	33.03%
	0.3	70.88%	76.55%	29.12%	23.45%	72.79%	27.21%
	0.4	81.05%	63.45%	18.95%	36.55%	75.11%	24.89%
	0.5	84.91%	56.55%	15.09%	43.45%	75.34%	24.66%
	0.6	90.53%	41.38%	9.47%	58.62%	73.95%	26.05%
	0.7	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.8	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.3.4 $\sqrt{X}$ 성장곡선 분석시 신경망 분석을 통한 간 질환 예측모형

$$y_k = g'\left(\sum_{j}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)}\left[\alpha_i + \beta_{i1} \times \sqrt{T}\right]\right)\right)$$

다음으로 $\sqrt{X}$함수를 이용한 성장곡선 분석 결과를 추정치로 사용한 신경망 모형에서는 iteration 6일 경우 Validation Set의 오차가 가장 최소를 나타내고 있다.

그림 13. $\sqrt{X}$성장곡선에 의한 신경망 모형의 Average Error Plot



이러한 신경망 모형의 Validation Set accuracy는 72.55%로 높게 나타나고 있으며 Sensitivity 역시 78.62%로 높은 값을 나타내고 있음을 표를 통해 살펴볼 수 있다.

표 29 $\sqrt{X}$성장곡선에 의한 신경망 모형의 Classification Table

	Predicted Probability	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	0.1	20.48%	96.99%	79.52%	3.01%	47.66%	52.34%
	0.2	56.02%	87.43%	43.98%	12.57%	67.18%	32.82%
	0.3	66.72%	83.06%	33.28%	16.94%	72.52%	27.48%
	0.4	74.55%	72.95%	25.45%	27.05%	73.98%	26.02%
	0.5	81.33%	66.39%	18.67%	33.61%	76.01%	23.99%
	0.6	87.80%	51.09%	12.20%	48.91%	74.75%	25.25%
	0.7	95.78%	27.87%	4.22%	72.13%	71.65%	28.35%
	0.8	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	0.9	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
	1.0	100.00%	0.00%	0.00%	100.00%	64.46%	35.54%
Validation Set	0.1	23.86%	95.17%	76.14%	4.83%	47.90%	52.10%
	0.2	59.30%	84.83%	40.70%	15.17%	67.90%	32.10%
	0.3	69.47%	78.62%	30.53%	21.38%	72.55%	27.45%
	0.4	78.60%	63.45%	21.40%	36.55%	73.48%	26.52%
	0.5	83.16%	54.48%	16.84%	45.52%	73.48%	26.52%
	0.6	88.42%	46.21%	11.58%	53.79%	74.18%	25.82%
	0.7	96.14%	24.83%	3.86%	75.17%	71.39%	28.61%
	0.8	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	0.9	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%
	1.0	100.00%	0.00%	0.00%	100.00%	66.27%	33.73%

6.4 모형 평가

표30. Cut-off Value 0.3일 경우 여러 가지 성장곡선과 예측 모형의 Classification Table

	모형	성장곡선	specificity	sensitivity	false negative	false positive	Total Accuracy	Miss Clasification
Training Set	로지스틱	$X_{(4)}$	65.51	83.06	34.49	16.94	71.74	28.26
		일차	61.45	81.42	38.55	18.58	68.54	31.46
		$LogX$	62.59	79.51	37.05	20.49	68.83	31.17
		$\sqrt{X}$	61.90	81.69	38.10	18.31	68.93	31.07
	나무모형	$X_{(4)}$	77.11	57.65	22.89	42.35	70.19	29.81
		일차	78.16	55.46	21.84	44.54	70.09	29.91
		$LogX$	70.78	78.42	29.22	21.58	73.49	26.51
		$\sqrt{X}$	80.27	64.75	19.73	35.25	74.75	25.25
	신경망	$X_{(4)}$	70.33	80.33	29.67	19.67	73.88	26.12
		일차	67.47	80.87	32.53	19.13	72.23	27.77
		$LogX$	69.13	78.42	30.87	21.58	72.42	27.58
		$\sqrt{X}$	66.72	83.06	33.28	16.94	72.52	27.48
Validation Set	로지스틱	$X_{(4)}$	68.77	74.48	31.23	25.52	70.69	29.31
		일차	62.46	81.38	37.54	18.62	68.83	31.17
		$LogX$	61.40	82.07	38.60	17.93	68.37	31.63
		$\sqrt{X}$	62.46	79.31	37.54	20.69	68.13	31.87
	나무모형	$X_{(4)}$	76.14	53.79	23.86	46.21	68.60	31.40
		일차	76.84	53.10	23.16	46.90	68.83	31.17
		$LogX$	70.53	67.59	29.47	32.41	69.53	30.47
		$\sqrt{X}$	76.14	60.69	23.89	39.31	70.93	29.07
	신경망	$X_{(4)}$	72.28	71.72	27.72	28.28	72.09	27.91
		일차	71.58	75.86	28.42	24.14	73.02	26.98
		$LogX$	70.88	76.55	29.12	23.45	72.79	27.21
		$\sqrt{X}$	69.47	78.62	30.53	21.38	72.55	27.45

Cut-off value 0.3 수준에서 각 성장곡선을 이용한 추정치를 통한 간 질환 예측모형들의 Classification Table을 통해 각 모형에 대한 평가를 실시하고자 한다. 먼저 Validation Set에 나타난 Accuracy와 Sensitivity를 살펴보면 로지스틱 회귀모형과 나무 모형의 Accuracy의 경우 약 68~70% 정도의 값이 나오는 반면 신경망 모형에서는 72%~73%정도의 값이 나오고 있다.

우선적으로 신경망 모형의 간 질환 예측률이 로지스틱과 나무 모형에 비해 우

수하게 나타나고 있음을 확인해 볼 수 있었다. 다음으로 간 질환을 예측하는 확률인 Sensitivity의 경우 로지스틱 모형이나 나무 모형, 신경망 모형간에 다소의 차이가 나타나고 있지만 결과를 살펴보면 대체적으로 나무 모형의 Sensitivity가 떨어지게 나타나고 있을 뿐 로지스틱 모형이나 신경망 모형간에는 별 다른 차이를 나타내고 있지 않고 있음을 확인해 볼 수 있다. 다만 성장곡선 분석을 통해 추정된 값들 간에 차이를 나타내고 있다. 먼저 로지스틱 모형과 같은 경우 Validation Set에서 4번째 시점에서의 건강검진 자료로 추정한 값을 사용한 경우 Sensitivity가 74.48%로 일차(81.38%), $LogX$ (82.07%) 또는 $\sqrt{X}$(79.31%)를 통해 추정된 값보다 낮게 나타나고 있음을 알 수 있다. 더 나아가 나무모형이나 신경망 모형에서도 이와 같은 양상을 나타내고 있는데 신경망 모형의 경우 4번째 건강검진 자료로 추정한 값을 사용한 경우 Sensitivity가 71.72를 나타내고 있는 반면에 일차 선형모형을 통한 성장곡선의 추정치를 사용할 경우 75.86%, $LogX$의 경우는 76.55%, $\sqrt{X}$의 경우는 78.62%를 나타내고 있음을 확인해 볼 수 있었다. 이와 같은 결과를 가져온 원인에 대한 분석을 하면 건강한 사람이 간 질환에 걸릴 경우 간 질환에 영향을 주는 Risk Factor의 변화에 대해 생각해 볼 수 있다. 아직 건강검진을 받지 않았지만 미래 시점에서 간 질환에 걸릴 경우에 대해 관심을 가질 경우 가장 마지막에 건강검진을 받은 자료에 기초하여 간 질환을 예측하게 되는데 이러한 경우 Risk Factor의 변화량을 고려치 못하게 된다. 성장곡선을 통해 미래시점에서의 Risk Factor의 유의적인 변화량을 통해 예측할 경우는 전체 Accuracy에 대해 크게 향상시키지는 못했지만 간 질환을 예측할 수 있는 확률인 Sensitivity를 향상시키는 결과를 가져온 것이다. 간 질환에 대한 성장곡선과 여러 가지 판별 모형들을 통해 모형의 우수성을 평가해 볼 수 있었는데 모형의 Accuracy와 같은 경우는 판별모형의 형태에 따라 많은 영향을 받은 반면 Sensitivity와 같은 경우에 대해서는 영향을 줄 수 없었고 다만 성장곡선을 통해 추정된 값을 사용할 경우 더욱 우수한 값의 Sensitivity의 값을 가질 수 있었다.

제 7 장 토의 및 결론

건강검진 자료(Screening Test Data)는 현재의 건강상태를 알아보거나 질병에 대한 조기진단이라는 두 가지 기능이 있다. 건강검진 자료 분석을 통해 질병에 대한 조기진단이라는 측면에서 간 질환에 관해 연구를 하였으며 이 과정 중 새로운 정보를 많이 얻을 수 있었다. 첫 번째로 간 질환에 관련하여 건강검진을 많이 받은 사람일수록 간 질환 발병률이 낮다는 것이다. 즉 자신의 건강에 대해 관심을 가지고 규칙적으로 검진을 받을 경우 간 질환에 대해 조기에 예방할 수 있다는 것이다. 두 번째로 간 질환에 관한 Risk Factor의 경우 일반적으로 알려진 간 기능 수치와, 음주력, 흡연력, 연령, 성별 외에도 종사직종이나 당뇨병력 등이 중요한 Risk Factor가 될 수 있으며 더 나아가 신체 모든 기능과 상태가 간 질환의 Risk Factor가 될 수 있다는 점이다. 즉 간 질환이라고 해서 간과 관련된 항목의 수치만이 중요한 것이 아니라 폐 기능 수치, 혈청지질, 갑상선 기능, 대사 및 전해질 수치 등 신체와 관련된 모든 항목이 Risk Factor가 될 수 있다는 점이다.

간 질환에 영향을 주는 Risk Factor의 규명을 위해 판별 분석을 실시했다면 앞으로 건강검진을 받을 경우(미래시점) 간 질환이 발생을 예측하는 모형의 추정을 위해 성장곡선분석을 실시하였다. 다양한 성장곡선 모형들과 판별 모형들을 통해 간 질환 예측모형을 추정한 결과 간기능 수치와 신체계측 지수인 BMI 수치를 중심으로 모든 건강검진 항목이 중요한 Risk Factor가 될 수 있음을 재확인하였다.

성장곡선 분석을 통해 간 질환을 간 질환이라고 예측할 수 있는 Sensitivity값을 많이 향상시킬 수 있었지만 전체 accuracy의 값에는 크게 변동이 없었음을 확인해 볼 수 있었다. 즉 성장곡선 분석에서 유의적인 기울기를 가진 성장곡선만을 선택하는 과정에서 나타난 결과라고 생각된다. 유의적인 기울기를 가진 성장곡선들은 간 질환과 관련된 Risk Factor가 시간의 흐름에 따라 일정수준을 계속 유지하기 보다는 상승과 하락과 같은 변동을 나타나고 있다는 의미이다. 간 질환과 관련된 Risk Factor의 변화는 미래시점에서의 간 질환에 걸릴 예측 모형의 추정에서 Specificity 값의 상승의 주요 원인이 되었던 것이다.

성장곡선 분석시 일차모형, 이차모형, Log X 모형, $\sqrt{X}$모형을 사용하여 Risk Factor의 미래 시점 값을 예측하였는데 이들은 모두 Linear한 형태의 모형들이다. 이 모형들을 토대로 non-linear한 모형으로 발전시켜 분석을 할 경우 더 많은 Specificity값의 상승을 기대할 수 있으며 accuracy의 상승 역시 기대할 수 있을 것으로 생각되어진다.

참 고 문 헌

강근석, 김충락, 회귀분석, 교우사, 1999

강현철, 한상태, 최종후, 김차용, 김은석, 김미경, 데이터마이닝 방법론 및 활용, 자유아카데미, 1999

강현철, 한상태, 최종후, 김은석, 김미경, 데이터마이닝 기능과 사용법, 자유아카데미, 2000

박종우, 현대역학(現代疫學), 연세대학교 출판부, 1999

백운붕, SAS 일반선형모형 분석, 자유아카데미, 1987

성웅현, 이승천, 회귀분석: 이론, 방법론, SAS 활용, 법문사, 2001

조용준, 데이터마이닝을 이용한 신경망 분석, SPSS 아카데미, 1999

Bishop, Neural networks for pattern recognition, Oxford Press, 1995

Curtis L. Meinert, Susan Tonascia, Clinical Trials Design, Conduct and Analysis, Oxford Press, 1986

Joseph P. Bigus, Data Mining With Neural Networks, McGraw-Hill, 1996

Kshirsagar, Anante M., Growth Curve, New York, 1995

M.Dossing, K.T.Petersen, M.Vyberg, J.H.Olsen , Liver Cancer Among Employees in Denmark, American Joural Of Industrial Medicine, 1997 ;32: 248-254

Neter, Kutner, Nachtsheim, Wasserman, Applied Linear Regression Model, IRWIN, 1996

Singer, Judith D., Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models, Journal of Educational and Behavioral Statistics, 1988; 23(4), 323 -355.

Stefano Porru, Donatella Placidi, Angela Carta, Umberto Gelatte, Maria Lusia Ribero, Alessandro Tagger,Paolo Boffeta, Francesco Donato, Primary Liver Cancer And Occupation In Men : A Case-Control Study In A

High-Incidence Area In Northern Italy, Publication of the International Union Against Cancer, 2001; 94:878-883

W.Y. LAU, Primary Liver Tumors, Seminars in Surgical Oncology, 2000;19:135-144

Yoshihisa Fujino, Tetsuya Mizoue, Noritaka Tokui, Takesumi Yoshimura, Prospective study of diabetes mellitus and liver cancer in Japan, Diabetes Metab Res Rev, 2001 ; 374-379

ABSTRACT

Screening Test Data Analysis for Liver Disease Prediction Model using Growth Curve

Kim, Young Sun
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

This study was researched based on screening test data accumulated from 1994 to 2001 for studying of risk factor related with liver disease and prediction model of liver disease.

In the existing study related with liver, the main current is the studing on liver cancer, not on liver disease, previous step into liver cancer.

As a result of estimating prediction model through the risk factors of liver disease and the growth curve on the basis of data, it is shown that most of the risk factors about liver disease are also those about well as liver cancer.

Moreover all the items in screening test have the possibility to become risk factors.

By examining frequency of screening test and liver disease prevalence rate, we can conclude that the more a man takse screening test, the lower the prevalence rate is.

The shows that it can be the prevention of liver disease to pay attention to

the health steadily and take screening test regularly.

In addition, to investigate liver disease prevalence from the viewpoint of the future, this study presumed risk factor through the various growth curve analysis and examined logistic regression, descion tree and neural network from those estimators.

In the case of neural network using growth curve estimator of $\hat{X}_{i(5)} = \alpha_i + \sqrt{\beta_i} T + \varepsilon_{iT}$, accuracy of liver disease was 72.55% and sensitivity was 78.62%. On the other hand in the case of liver disease prediction model using recent screening test data estimator, accuracy was 72.09% and sensitivity was 71.72%. Those are lower than liver disease prediction model of growth curve analysis.

In the various liver disease prediction models assumed by growth curve and many distinction models, when growth curve estimator was used, sensitivity value was improved.

According to discriminant model (logistic regression, decsion tree and neural network), accuracy made a difference.

Key Words : Growth Curve, Screening Test, Logistic Regression, Descion Tree, Neural Network, Accuracy, Sensitivity, Specificity, Liver Disease, Risk Factor