# 혼합 로지스틱 분포를 이용한 백내장 발생 예측 모형 연구

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

김 소 연

# 혼합 로지스틱 분포를 이용한
# 백내장 발생 예측 모형 연구

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2005년 12월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 소 연

# 김소연의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2005년 12월 일

# 목 차

# 표 차례



# 그 림 차 례

# 국 문 요 약

## 혼합 로지스틱 분포를 이용한 백내장 발생 예측 모형 연구

백내장은 시력장애를 초래하므로 건강검진 자료를 이용하여 조기진단이 이루어진다면 발생 위험을 크게 줄일 수 있는 질병이다. 본 논문에서는 1994년부터 2005년까지 건강검진센터에서 건강검진을 받은 126,532명 중 S병원에 내원한 결측치를 제외한 4,591명의 검진자를 연구대상으로 백내장 진단을 받은 검진자의 건강검진 자료을 토대로 백내장 발생에 대한 위험인자를 살펴보고 아울러 백내장 발생 예측을 위한 통계학적 모형을 구축하였다. 백내장 발생이 여러 가지 복합적인 위험요인에 의해 발생되는 질병이므로 그 특성에 따라 여러 개의 하위집단으로 이루어져 있다는 가정 하에서 혼합 로지스틱 분포를 이용하였고 기존의 데이터 마이닝 기법인 판별분석과 그 성능을 비교, 분석하였다.

본 연구에서는 독립변수의 개수에 따라 집단이 1~2개인 혼합 로지스틱 회귀모형을 적용하였는데, EM 알고리즘을 이용하여 최대우도 추정량을 구하였다. 추정된 모수를 이용하여 집단의 수는 AIC 값으로 결정하였고, 실제 자료에 어느 정도 가깝게 예측하는지를 데비언스 값으로 살펴보았다. 그 결과 하나의 로지스틱 회귀모형 보다는 집단이 두 개인 혼합 로지스틱 회귀모형을 적용했을 때 위험요인의 복합적인 영향에 대해 설명하고 있었고, 실제 백내장 발생에 대한 예측력이 높다는 것을 알 수 있었다. 또한, 선형판별분석과 이차판별분석을 통해 백내장 발생을 예측하여, 혼합로지스틱 회귀모형이 정확도 64.10%, 민감도 21.88% 으로 다른 기법에 비해 예측력이 높음을 확인하였다.

---

핵심이 되는 말 : 건강검진 자료, 위험요인, 백내장, 혼합 로지스틱 모형, 예측모형

# 제 1장 서론

백내장이란 눈에 있는 투명한 수정체에 어떠한 원인에 의해서 뿌옇게 혼탁이 발생한 상태를 말한다. 이로 인해 보고자하는 물체의 상이 수정체를 통과하지 못하고 망막에 정확하게 초점을 맺지 못함으로서 우리 눈의 가장 중요한 기능인 시력에 장애를 초래하는 질환이다. 이러한 백내장은 한국을 포함하여 세계적으로 실명을 초래하는 가장 큰 원인이며, 전 세계 실명자 3,500만 ~ 4,500만명 중 백내장이 그 절반인 2,000만명 가량을 차지하고 있으며, 평균 수명의 증가와 더불어 점점 증가하는 추세에 있다(백내장・녹내장 이상욱). 1990년의 WHO 보고에 의하면, 52억 전 세계 인구 중 백내장으로 실명한 사람이 약 1,300만명이라고 하며, 중국・인도・사우디아라비아・나이제리아・한국 등에서는 백내장이 실명원인 중 1위로, 미국은 3위로 보고가 될 정도로 백내장은 실명과 높은 관계가 있다.

우리나라의 경우도 산업발전과 경제성장에 따른 영양개선과 건강에 대한 관심의 증대, 의학기술의 발달 등에 힘입어 노령인구가 차지하는 비중이 커질 전망이다. 이에 따라 다른 연령층에서와는 다른 임상양상과 분포를 보이는 노화성 질환이 증가하는 추세를 보이고, 그 중 시력장애를 초래하는 질환은 그 관리의 중요성이 매우 커지고 있다. 이에 한국실명예방재단은 지난 1973년 창립직후부터 무의촌 지역 안검진 및 개안 수술비 지원사업을 매년 시행해 왔으며, 2003년부터 보건복지부의 예산 지원을 받아 전국 안과 무의촌 및 취약지역의 65세 이상 노인을 대상으로 정밀 안검진을 실시하고 백내장, 녹내장 등 안과 수술이 필요한 기초생활보장수급권자 및 저소득층에게 수술비를 지원하는 사업을 펼치고 있다. 2003년 전국 53개 지역 65세 이상 노인 7,750명을 대상으로 실시된 한국실명예방재단의 '2003년 노인 안검진 및 개안수술사업 결과'에 따르면 전체 검진 노인 가운데 7,168명(91.4%)이 안질환을 가지고 있었다. 그 중 4,383명(45.4%)이 백내장으로 가장 많아 치료 대책이 시급함을 알 수 있다.

이와 같이 백내장은 국가경제와 문화발달로 노령인구가 증가하고 있는 시점에서, 사회, 경제적으로 심각한 문제임을 알 수 있다. 그럼에도 불구하고 국내외에

서 백내장 발병 전 예후에 대한 연구와 백내장의 위험요인에 대한 연구는 제한적으로 이루어져 있다. 이러한 백내장 발생에 대한 조기 예측 및 관리가 이루어진다면 발병률을 충분히 줄일 수 있는 질병이 될 것이다.

본 논문에서는 1994년 5월부터 2005년 9월까지의 건강검진센터에서 건강검진을 받은 126,532명 중 S병원에 내원하여 안과 정밀 검사를 받은 검진자 5,804명을 연구대상으로 하였다. 백내장 진단을 받은 검진자의 건강검진 자료(screening data)을 토대로 백내장 발생에 대한 위험인자(risk factor)를 살펴보고 아울려 백내장 발생 예측을 위한 통계학적 모형을 구축하였다. 건강검진 자료들이 여러 개의 집단으로 이루어져 있다는 가정 하에서 혼합 로지스틱분포(Logistic Mixture)을 이용하여 백내장 발생에 대한 예측모형을 설정하였다. 또한 기존의 데이터마이닝 (data mining)기법인 판별분석(discriminant analysis), 로지스틱 회귀분석(logistic regression)과 그 성능을 비교 분석하였다.

# 제 2장 여러 가지 판별 모형

본 논문에서는 백내장 발생을 건강검진 자료을 통해서 예측할 수 있는 통계적인 모형을 제시하고자 한다. 검진자의 기초정보, 신체계측, 대사 및 전해질 검사, 간기능 검사, 신 기능 검사, 소변 검사 등으로부터 각 검진자의 백내장 위험 정도를 얼마나 정확하게 예측하느냐에 중점을 두고 있다.

백내장 발생을 분류, 예측하는 방법에는 혼합 로지스틱 회귀분석 (Mixtures of Logistic regressions)을 사용하였으며 이 방법과 성능을 비교하기 위한 분류기법은 다음과 같다. 가장 보편적으로 많이 사용하는 선형판별분석, 이차판별분석, 로지스틱 회귀분석이 사용되었다.

본 장에서는 혼합 로지스틱 회귀분석의 이론을 살펴보기 앞서 여러 가지 분류기법인 선형판별분석, 이차판별분석, 로지스틱 회귀분석의 이론을 살펴보기로 한다.

## 2.1 판별 분석

두 개 또는 그 이상의 모집단으로부터 얻어진 관찰값을 바탕으로 이들 모집단을 가장 잘 분리할 수 있는 어떤 판별기준 또는 판별함수를 구한 후 이것을 이용하여 기존의 관찰값을 소속 모집단에 할당하는 통계적 방법을 판별분석이라 한다. 이러한 절차에 의해 분류된 분류표로부터 판별함수의 오분류율을 평가하고 또한 소속 집단이 알려지지 않은 새로운 개체를 어떤 모집단에 분류하는 과정을 거치게 된다. 그러므로 판별분석은 판별함수를 구하고 구해진 판별함수에 따라 기존의 개체를 할당할 뿐만 아니라 미지의 새로운 개체를 특정 모집단에 분류하는데 더 큰 의의를 가질 수 있다.

관찰값들이 다변량 정규 분포를 따르는 모집단의 경우, 분류방법은  두 집단의 공분산 행렬의 동일성 여부에 따라 크게 두 가지로 나눌 수 있다. 두 집단의 공분산 행렬이 같을 시에는 선형판별분석 (Linear Discriminant analysis ; 이하 LDA),

두 집단의 공분산 행렬이 다른 경우의 분류 방법은 이차 분류 방법 (Quadratic Discriminant analysis; 이하 QDA) 이다

### 2.1.1 선형판별분석

각 집단의 확률밀도 함수 $f_i(x)$는 평균벡터 $\mu_i$ 이고 공분산행렬이 $\Sigma_i$ 인 다변량 정규밀도함수라고 가정하자. 두 집단의 공분산 행렬이 같은 경우 ($\Sigma_1 = \Sigma_2 = \Sigma$ ), 두 확률밀도함수의 비율에 자연대수를 취하면

$$L(x) = \ln\left\{\frac{f_1(x)}{f_2(x)}\right\} = \ln f_1(x) - \ln f_2(x)$$

$$= (\mu_1 - \mu_2)^T \Sigma^{-1} x - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2)$$

이다. 관측값 $x$에 대하여 선형이므로 $L(x)$를 모집단 선형분류함수라고 한다. 여기서 $L(x)$는 두 집단을 구분하는 변수들의 선형 결합식이므로 모집단 선형판별 함수라고도 부른다. 모집단의 모수 $\mu_1$, $\mu_2$, $\Sigma$ 들이 일반적으로 알려져 있지 않기 때문에 표본으로부터 추정해야 한다. 각 집단의 표본평균벡터를 $\bar{x}_i$ , 합동 표본 공분산 행렬을 $S_p$ 라고 할 때의 선형판별 함수는

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

으로 정의한다. $\pi_i$ 를 각 집단의 사전확률이라고 할 때, 위의 판별 함수 값이 다음 조건을 만족하면 새로운 관측값 $x_0$는 집단 1에 분류하고, 그 이외 경우에는 집단 2로 분류하면 다음의 식으로

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

표현한다.

### 2.1.2 이차선형판별분석

두 집단의 공분산행렬이 다른 경우 분류함수는 공분산행렬이 같은 경우보다 약간 복잡한 형태가 되고 그 형태는

$$Q(x) = \ln \left\{ \frac{f_1(x)}{f_2(x)} \right\} = \ln f_1(x) - \ln f_2(x)$$

$$= -\frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}) x - k$$

이다. 여기서 $k = \frac{1}{2} \ln (|\Sigma_1| / \Sigma_2) + \frac{1}{2} (\mu_1^T \Sigma_1^{-1} \mu_1^T - \mu_2^T \Sigma_2^{-1} \mu_2^T)$이다. 따라서 두 집단의 공분산행렬이 다르면 분류함수의 우변 첫째 항이 관측값 $x$의 이차형식으로 표시되므로 $Q(x)$를 모집단이차분류함수 혹은 모집단이차판별함수라고 부른다. 두 집단의 모수가 알려져 있지 않은 경우에는 선형판별분석과 같이 표본으로부터 추정해야 한다. 각 집단의 표본평균벡터를 $\overline{x}_i$ , 표본 공분산 행렬을 $S_i$ 라고 할 때의 이차판별 함수는

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

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

$\pi_i$ 를 각 집단의 사전확률이라고 할 때, 위의 판별 함수 값이 다음 조건을 만족하면 새로운 관측값 $x_0$는 집단 1에 분류하고, 그 이외 경우에는 집단 2로 분류하면 다음의 식으로

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

표현한다.

## 2.2 로지스틱 회귀분석

종속변수가 두 가지 값만 취하는 질적인 이분형 변수일 때 이때의 종속변수 $y_j$가 1 혹은 0의 값을 가진다면 베르누이 확률변수가 된다. 종속변수가 1을 가질 경우을 성공할 확률 $\theta_j$ 라고 정의하고 $\theta_j$ 에 로짓 변환(logit transformation)을 하였을 경우 다음과 같이

$$logit(\theta_j) = ln\frac{\theta_j}{1-\theta_j} = \beta^T x_j \qquad (j = 1, 2, ..., p)$$

표현한다. 이것을 로지스틱 반응함수라 부르고 위 식을 다시 표현하면

$$\theta_j = \frac{\exp(\beta^T x_j)}{1+\exp(\beta^T x_j)} = \frac{1}{1+\exp(-\beta^T x_j)}$$

이다. $p$개의 독립변수 $x_p$ 에 따른 관찰치 $y$가 1로 분류될 확률인 $P(Y=1|x_{1,}x_2,\ldots,x_p)$ 에 대해서 정리하면

$$P(Y=1|x_{1,}x_2,\ldots,x_p)=\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ \cdots\ +\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ \cdots\ +\beta_px_p)}$$

을 따르게 된다. 로지스틱회귀모형에서 모수인 회귀계수 $(\beta_0,\beta_1,\ \cdots,\beta_p)$를 추정하는데 가장 널리 사용되는 방법은 최대우도방법 (maximum likelihood method)이다. 최대우도추정은 모집단에서 얻어진 표본들이 독립이라는 가정 하에 로그우도함수의 최대화로부터 얻어진다. 로그우도함수는 다음의

$$\ln[L(P;y)]=\sum_{j=1}^{n}y_j\ln\left[\frac{P(x_j)}{1-P(x_j)}\right]+\sum_{i=1}^{n}\ln[1-P(x_j)]$$

과 같다. 위 식을 최대로 하는 회귀계수의 최대우도 추정값은 선형이 아니므로 직접 구할 수 없고, 뉴턴 - 랩슨 (Newton-Raphson) 방법 혹은 피셔(Fisher)의 스코어링방법 (method of scoring)등과 같은 반복적인 (iterative) 추정 방법에 의하여 근사해를 구할 수 밖에 없다. 회귀계수의 최대우도 추정값이 구해지면, 각 개체별로 어떤 사건이 발생될 확률인 사후확률 $\hat{P}_x=P(Y=1|x_{1,}x_2,\ldots,x_p)$은 아래와 같이

$$\hat{P}_x=\frac{\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ \cdots\ +\hat{\beta}_px_p)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ \cdots\ +\hat{\beta}_px_p)}$$

표현하며, 이것을 이용하여 추정할 수 있다. 이렇게 얻어진 각 개체에 대한 사후확률은 그 개체를 분류하기 위해 사용될 수 있다. 즉 추정된 사후확률은 0 과 1 사이의 값을 가지게 되므로, 적절한 경계값(cutoff value), $c$ 을 정하여 이 값을 기

준으로 각 개체를 다음과 같이

$$\begin{cases} \hat{p}_x \geq c \text{ 이면, 집단 1로 분류} \\ \hat{p}_x < c \text{ 이면, 집단 0으로 분류} \end{cases}$$

으로 분류하는 것이다. 로지스틱 회귀방법은 분류기법에서 가장 널리 사용되는 기법으로 선형성의 가정으로 회귀계수나 오즈비(odds ratio)를 통해 해석이 용이하며 많은 정보를 제공해 준다. 그러나 비선형성의 자료인 경우는 예측의 한계성을 가진다.

# 제 3 장 혼합 로지스틱 회귀모형

## 3.1 모형의 설정

$p$ 개의 독립변수 $X_1, X_2, \cdots, X_P$ 에 따라 자료가 각 성분(component) 혹은 군(group)으로 이루어져 있다고 가정한다면 혼합모형(mixture model)을 고려해 볼 수 있다. 이는 다음과 같이

$$f(y_j) = \sum_{i=1}^{g} \pi_i f_i(y_j;\theta_{ij}) \quad ,$$

$$\sum_{i=1}^{g} \pi_i = 1 \quad , \ \pi_i \geq 0 \quad , \ j = 1, ..., N$$

으로 정의한다. 여기서 $G$는 자료의 성분 개수를 의미하고 $\pi_i$는 $i$번째 성분에 속할 혼합 비율 (mixing proportion)을 나타내고, $f_i$ , $\theta_{ij}$ 는 혼합모형(mixture)에서 $i$번째 성분의 밀도함수 (density)와 모수들을 의미한다. 이 때, 독립적으로 관찰된 종속변수 $y_j$ 가 0 혹은 1을 갖는 이분형일 경우 각 성분의 밀도함수는 이항 분포(binomial distribution)를 따르며 다음과 같이

$$f_i(y_j;\theta_{ij}) = \binom{N_j}{y_j} \theta_{ij}^{y_j} (1-\theta_{ij})^{N_j - y_j} I_{A_j}(y_j) \ , \ A_j (= 0, 1, ..., N_j)$$

정의된다(McLachlan et al 2002). 즉, 종속변수 $y_j$ 은 $N_j$ 개 중 성공한 횟수에 해당되며 $\theta_{ij}$ 는 베르누이 시행에서 성공할 확률을 말한다. 성공할 확률 $\theta_{ij}$ 에 로짓 변환을 하면 로지스틱 반응함수가 되며 다음의

$$logit\,(\theta_{ij}) \;= ln\frac{\theta_{ij}}{1-\theta_{ij}} \;= \beta_i^T x_j \quad (i=1,\,\cdots,g\,;j=1,\,\cdots,n\,)$$

과 같다. 따라서 종속변수 $y_j$ 가 이분형일 경우의 혼합모형은 혼합 로지스틱모형임을 알 수 있다. 혼합비율 $\pi_i$는 독립변수로 이루어진 로지스틱 반응함수를 가지고, $j$번째 관측치 $y_j$ 가 $i$번째 성분에 속할 혼합 비율은

$$\begin{aligned}\pi_{ij} &= \pi_i\,(x_j;\alpha\,)\\ &= \exp\,(\omega_i^T x_j)/\{1+\sum_{h=1}^{g-1}\exp\,(\omega_h^T x_j)\,\}\;\;(i=1,\,....,\,g\,)\end{aligned}$$

으로 정의되며 , 여기서 $\omega_g = 0,\alpha = (\omega_1^T,\,....,\,\omega_{g-1}^T\,)^T$ 으로 로지스틱 회귀계수에 해당이 된다.

혼합 로지스틱 회귀모형의 평균(mean)과 분산(variance)은 다음과 같이

$$E(\,Y_j) = \sum_{i=1}^{g}\pi_{ij}\theta_{ij}$$

$$var\,(\,Y_j) \;=\; N_j\,(\,\sum_{i=1}^{g}\theta_{ij}\,)\,(1-\sum_{i=1}^{g}\theta_{ij}\,)$$

표현된다. 혼합 로지스틱회귀 모형의 모수들을 추정하기 위해서는 앞장의 로지스틱 회귀모형과 같이 최대우도 추정법을 사용한다. 이 때, 혼합 로지스틱 회귀 모형의 우도 함수는 $n$개의 자료에 대한 결합 확률함수 (joint probability function)로 표현 할 수 있고 다음과 같이

$$L = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_{ij} f(y_j;\theta_{ij})$$

정의된다. 혼합 로지스틱 회귀 모형의 모수 벡터를 $\Psi = (\pi_{ij}, \theta_{ij})^T = (\alpha^T, \beta^T)$로 정의를 하면, $\Psi$ 를 찾기 위하여 우도함수 $L$ 에 로그를 취한 $l = \ln L$ 을 최대화함으로서 모수들을 추정할 수 있다. 즉, 로그우도함수는 다음과 같이

$$\log L = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_{ij} f(y_j;\theta_{ij})$$

표현한다.

## 3.2. 모형 추정을 위한 EM 알고리즘

우도함수에 관한 추정 시 모수들이 비선형 함수를 가질 경우에는 직접 구할 수 없으므로 수치적 최적화 기법을 고려하는데, 많은 경우에서 사용되어 지는 것은 EM (Expectation - Maximization) 알고리즘이다 (Baum et al.,1970 ; Dempster et al., 1977). EM 알고리즘은 관측치 외에 다른 변수를 결측치 (missing value) 또는 잠재치 (latent value)로 생각하여 실제 관찰된 변수에 포함시킨 후, 이를 통해 얻어지는 단순화된 문제를 갱신해가면서 해를 구하는 방법이다.

관찰된 자료 $y$ 가 주어졌을 때, '결측치' 혹은 '잠재치'라고 불리우는 자료를 $y_{mis}$라고 하자. 그러면 $y$ 는 불완전자료 (incomplete data)라고 하고, $(y, y_{mis})$는 완전자료(complete data)가 된다. 완전 자료에 대한 집합 $L = \{y, y_{mis}\}$에 대하여 최대우도 추정법을 고려한다 (Little and Rubin , 1987).

완전자료의 로그 우도함수에 기초하여 모수들을 갱신 할 때 주의할 점은 완전

자료에 대한 로그 우도함수를 바로 구할 수 없다는 점이다. 이는 이러한 우도함수는 불완전 확률변수의 함수이기 때문이다. 그렇다면, 완전자료에 대한 로그우도함수의 기대값 $E_{y_{mis}}[\ln P(y, y_{mis}|\Psi)]$를 최대화하는 $y_{mis}$를 구하는 방법을 생각할 수 있다. 이러한 방법이 EM 알고리즘의 근간이 된다. EM 알고리즘의 목적은 완전자료의 로그 우도함수의 기대값을 최대로 하는 모수값을 반복알고리즘을 통하여 관측값으로부터 만들어지는 로그 우도함수를 최대로 하는 모수값으로 찾아가는데 있다.

혼합 로지스틱 회귀모형에 EM 알고리즘을 적용하기 위하여 성분에 대한 가변수 (indicator variable) $z_{ij}$를 정의할 수 있다. 즉, $j$번째 종속변수가 $i$번째 성분에 속하는 여부에 따라 0 혹은 1을 갖는 변수이고 다음과 같이

$$z_{ij} = \begin{cases} 1 & ; \text{ 집단 } i \text{에 속한 경우} \\ 0 & ; \text{ 집단 } i \text{에 속하지 않은 경우} \end{cases}$$

표현한다. 관찰되지 않은 $z_{ij}$가 알려져 있다면 완전 자료의 로그 우도함수는

$$\log L_c(\Psi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\left[\log \pi_{ij} + log f_i(y_j;\theta_{ij})\right]$$

으로 정의된다. 독립변수 $x_j$ 을 가지는 종속변수 $y_j$가 주어졌을 때, $z_{ij}$의 조건부 기대값은 다음의

$$\begin{aligned} \tau_i(y_j;\Psi) &= P(Z_{ij}=1|y_j, x_j) \\ &= \frac{\pi_i f_i(y_j|x_j,\theta_i)}{f(y_j|x_j,\Psi)} = \frac{\pi_i f_i(y_j|x_j,\theta_i)}{\sum_{h=1}^{g}\pi_h f_h(y_j|x_j,\theta_h)} \quad , \quad h=1,2,\cdots,g \end{aligned}$$

과 같다.

EM 알고리즘은 로그함수의 기대값을 구하는 E-step과 기대값의 최대값을 구하는 M-step으로 구성된 알고리즘으로 반복함으로서 $\hat{\Psi}$을 계산한다. k번의 반복을 가정한다면 모수의 추정값들 $\Psi^{(k)} = (\pi_{ij}^{(k)}, \theta_{ij}^{(k)})^T = (\alpha^{(k)^T}, \beta^{(k)^T})^T$ 이다.

E-step :

$$\tau_i = P(Z_{ij} = 1 | y_j, \Psi^{(k)}) = \frac{\pi_i \, f_i \, (y_j \, ; \theta_{ij}^{(k)})}{\sum_{h=1}^{g} \pi_h f_h \, (y_j \, ; \theta_{hj}^{(k)})}$$

이 단계에서는 관찰치들이 각각의 집단 분포에 속하는지를 추정하는 단계이며, 관찰치 $y_j$가 집단 $i$에 속하는 사후 확률 (posterior probability)을 의미한다.

M-step :

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij} \, (y_j ; \Psi^{(k)}) \, \partial log \pi_{ij} / \partial \alpha = 0$$

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij} \, (y_j ; \Psi^{(k)}) \, \partial \log f_i \, (y_j ; \theta_{ij} \, ) / \partial \beta = 0$$

여기서 $\tau_{ij}(y_j, \Psi^{(k)}) = \frac{\pi_{ij} \, f_i \, (y_j \, ; \theta_{ij}^{(k)})}{\sum_{h=1}^{g} \pi_{hj} f_h \, (y_j \, ; \theta_{hj}^{(k)})}$ 이다.

이 단계에서 사후확률을 가중치(weight)로 이용하여 수렴할 때까지 반복하는 IRLS (Iteratively re-weighted least squares)방법을 사용한다. EM 알고리즘으로 모수를 추정하는 과정에서 국소 최대값(local maximum)이 발견 될 수 있으므로

다양한 초기값들을 가지고 여러 번 계산을 해야 하며, 가장 큰 로그 우도함수를 가지는 추정값으로 선택하는 것이 바람직하다.

## 3.3 집단 수의 결정

우리가 알지 못하는 모수들을 EM 알고리즘을 통해 가장 큰 로그 우도함수를 갖는 모수값을 찾는다. 그 다음으로 생각할 것은 몇 개의 집단으로 나누어야 하는지, 즉 몇 개의 집단으로 이루어져 있는지를 결정해야한다. 집단의 수를 결정하는데 사용하는 방법으로 정보에 근거한 척도들이 있다. 첫 번째 척도는 가장 많이 사용되는 Akaike(1973)가 제안한 정보기준인 AIC (Akaike's Information Criterion)이고, 두 번째 척도는 Schwartz(1978)가 제안한 베이지안 정보기준인 BIC (Bayesian Information Criterion)이다. 이는 다음과 같이

$$AIC = -2Log\,L(\hat{\Psi}) + 2v_g$$

$$BIC = -2Log\,L(\hat{\Psi}) + v_g log(N) \qquad v_g : \text{모형에서 모수들의 수}$$

정의된다. 집단의 수인 $g$ 의 값을 1부터 다양하게 주어 혼합 로지스틱 회귀모형에 적용한 후에 AIC 또는 BIC에서 첫 번째 국소 최소값(local minimum)에 해당하는 모형을 선택한다.

## 3.4 적합도

혼합 로지스틱 회귀모형에 적용할 경우 종속변수가 1이 될 확률값이 나오게 된다. 이 확률값이 실제 종속변수에 얼마나 가까운지, 즉 설정된 모형이 자료를 얼마나 잘 적합시키고 있는지 살펴보아야 한다. 이는 설정된 모형과 연관된 오차를 반영해주는 데비언스(deviance)을 이용할 수 있다.

데비언스는 다음과 같이

$$D = -2Log\,L(\hat{\Psi})$$

표현한다. 데비언스 통계량 $D$는 로그 우도함수의 -2배를 해준 값으로 작을수록 적합도는 좋아지게 된다.

# 제 4장 건강검진 자료을 이용한 백내장 발생에 대한 연구 모형

## 4.1 건강검진 자료

### 4.1.1 소개

건강검진은 만성질환을 일으킬 수 있는 위험요소나 신체 변화를 조기발견, 이에 대처하며 건강한 삶을 유지할 수 있는 생활습관 등을 배우고 실천할 수 있는 기회를 갖게 해준다. 본 논문에서는 1994년 5월 30일 ~ 2005년 9월 30일 사이에 건강검진을 받은 126,532명의 검진자를 그 대상으로 백내장 발생에 대한 연구를 시작하였다.

### 4.1.2 건강검진 항목

건강검진 센터에서 검진하고 있는 기본 검사 항목들은 기초정보인 성별 및 연령 신체 계측 지수, 폐기능 검사, 혈액검사, 당뇨 · 신장기능 · 전해질 검사 , 간기능 검사, 안과검사, 청력검사, 심전도 검사, 영양상태 등이 있다. 이러한 항목 중 임상적으로 알려진 백내장의 위험요인은 기초정보, 신체 계측 지수, 당뇨 · 신장기능 · 전해질 검사, 간기능 검사 등이다. 이러한 검사항목들이 백내장 발생을 예측하는데 어느 정도의 위험인자로 존재하는지, 어떠한 영향을 미치는 지에 대해 살펴 볼 것이다. 아래의 표1는 검진센터에서 검진하는 항목들 중 세부사항에 대한 것이다.

표1. 건강검진 검사항목

| 검사항목 | 변수 | | 검사항목 | 변수 |
|---|---|---|---|---|
| 기초정보 | 성별 | | 대사 및 전해질 | 나트륨 |
| | 연령 | | | 칼륨 |
| 신체계측 | 신장 | | | 염소 |
| | 체중 | | | 이산화탄소 |
| | BMI | | | 칼슘 |
| 혈액검사 | 적혈구 | RBC | | 인 |
| | | Hb | | 혈당 Glucose |
| | | Hct | | 당화혈색소 |
| | | MCV | | 혈중요소질소 |
| | | MCH | | 크레아티닌 |
| | | MCHC | | 요산 |
| | 백혈구 | WBC | 혈청지질 | 총콜레스테롤 |
| | | 임파구 | | 중성지방 |
| | | 호산구 | | 고밀도콜레스테롤 |
| | 혈소판 | 혈소판 | | 저밀도콜레스테롤 |
| | 혈액형 | ABO | 뇨검사 | 비중 |
| | | RH | | 산도 |
| 간기능 | 총단백 | | | 단백 |
| | 알부민 | | | 요당 |
| | 총빌리루빈 | | | 케톤체 |
| | AST(GOT) | | | 잠혈 |
| | ALT(GPT) | | | 요빌리노겐 |
| | Alkaline phosphatase | | | 빌리루빈 |
| | r-GTP | | | 아질산염 |
| 안과 | 시력 | | | 백혈구 |
| | 안압 | | | 색 |
| | 안저 | | | 탁도 |

## 4.2 백내장 발생에 대한 위험요인

### 4.2.1. 연령

백내장은 나이가 들어가면서 발생하고 서서히 증가하기 때문에 연령 자체가 가장 큰 위험인자라 할 수 있다. 하지만 많은 연구에도 불구하고 어느 연령에서 수정체 혼탁이 일어나기 시작하는지 정확히 알려져 있지 않고 있다. 보통 40세가 지나면 서서히 수정체 혼탁이 발생하는 것으로 여겨지고 있다. 그러나 고령자라고 해서 모두 백내장이 발생하는 것은 아니고 백내장이 있다고 해도 모두 병적인 경우는 아니기 때문에 모든 노인백내장이 연령과 관련이 있다고 보기는 어렵다. 따라서 백내장의 발생은 여러 위험인자들이 관련된 복합적인 원인에 기인한다고 보아야 하겠다.

### 4.2.2. 성별

성별에 따른 여러 유병률 조사에서 여자의 백내장 유병률이 높게 나타나고 있다. 건강보험 심사평가원에서 1999년도 노인성 백내장 수술건수는 18만 8,595건이었고, 성별 노인성 백내장 수술건수는  남자 32%, 여자 68%이라고 발표하였다. 이것은 여자의 백내장 유병률이 높음을 뒷받침 해주고 있다. 그러나 여러 다른 연구에서  조사 당시 표본 수, 조사대상에 따른 차이가 있어, 성별과 백내장의 유병률 사이에 관련이 없다는 주장도 있다. Hiller는 세 기관의 자료를 이용하여 연령을 보정한 상대적 백내장 위험도를 조사했는데 여성의 유병률이 남성보다 13% 정도 높은 것으로 나타났다. 이러한 성별의 차이가 많다는 원인은 아직까지 밝혀지지 않고 있다.

### 4.2.3 전신질환

신체 내에서 발생한 대사 장애와 관계되며, 구체적인 발병으로는 당뇨병, 고혈압, 고지혈증, 갑상선 질환, 부갑산성 질환 등이 있다.

특히 Vogt, Mayerm Hiller 등은 당뇨병이 백내장의 발생과 밀접한 연관 관계가 있다고 한다. 신경환 등(1992)은 당뇨병 환자군에서 4 ~ 5배 가량 백내장 발생위험이 높다고 보고하였다. 당뇨병으로 인한 백내장 유발에는 몇 가지 기전이 제시되어 있다. 먼저, 혈당 증가 시 수정체 내 소르비톨(sorbitol)이 축적되어 삼투압이 증가함에 따라 수정체섬유뷰종이 일어나며 결국 혼탁을 나타내게 된다는 것이다. 둘째로, 수정체 단백질의 글라이코실화(glycosylation)으로 인해 산화가 쉽게 유발되어 수정체 혼탁을 일으킨다는 것이다.

그 밖에도 분만횟수와 백내장 발생과의 관련성을 조사한 결과 3회 이상의 분만 경험이 이는 군에서 2회 미만 분만군보다 백내장 유병률이 약간 높게 나타난다고 보고 되어 있다.

### 4.2.4 환경요인

(1) 알코올 섭취

알코올 섭취와 백내장 발생과의 연관관계에 대한 연구가 최근 널리 행해지고 있다. 그 중 한 연구에서 중등도 이상의 알코올 섭취가 백내장 발생을 증가시키며 주로 후낭하 백내장이 발생과 연관이 있다고 하였고, 다른 연구에서는 중등도 알코올 섭취군에서만 백내장 발생률이 낮은  U자형 연관관계가 있다고 하였다. 이는 알코올 자체의 직접적인 독성, 대사산물(아세트알데히드)에 의한 독성, 알코올 대사로 인한 탄수화물 대사 및 황산화제 농도 변화에 따른 간접적인 영향이 알코올 섭취로 인한 백내장 발생의 기전으로 알려져 있다.

(2) 일광노출

눈의 각막은 일광노출 정도에 따라 여러 조직에 손상을 받는 것으로 알려져 있다. 자외선의 파장에 따른 흡수정도에 따라  광각막염 및 결막염 등의 급성영향과 백내장과 같은 장기 영향을 가져올 수 있다. 일광 노출과 백내장 발생과의 관련성은 여러 역학 조사에서 보고하고 있는데 호주의 Hollow와 Moran 등은 64,307명의 원주민과 41,254명의 비원주민 백내장 환자들을 조사한 결과 자외선 노출 빈도가 높은 지역에서 백내장 유병률이 훨씬 높다고 보고하였다.

(3) 흡연

흡연은 백내장 발생빈도를 증가시키는 가장 잘 알려진 원인 중 하나이며, 주로 핵 백내장을 야기하며 피질백내장 빈도와는 연관성이 없다고 알려져 있다. 흡연이 백내장 발생을 증가시키는 정확한 기전은 알려져 있지 않지만, 수정체 내의 산화 작용 증가가 원인으로 보인다. 동일량의 카로티노이드 섭취군을 비교했을 때 흡연 환자군에 혈장 케타카로틴 농도가 비흡연군에 비해 감소되었다는 연구결과들이 이를 뒷받침한다.

### 4.2.5 유전적 요소

백내장이 가족력으로 발병하는 경우나 쌍둥이 형제에서 같은 증상의 백내장이 생기는 경우를 흔히 볼 수 있다. 그러나 아직까지도 유전적 소인에 관한 정확한 규명은 밝혀진 바가 없다.

### 4.2.6 영양상태

필수 아미노산, 칼슘, 비타민 등의 결핍으로 백내장이 발생할 수 있으며 특히 저개발국에서 중요한 원인이 되고 있다.

### 4.2.7 기타

안과 질환 중에는 합병증에 의해 백내장을 일으키기 쉬운 것들이 있는데 그 대표적인 것으로서 포도막염과 녹내장 등이 있다.

## 4.3 백내장 발생 예측 모형

### 4.3.1 분석자료

건강검진 센터에서 1994년 5월 30일 ~ 2005년 9월 30일 사이에 건강검진을 받은 126,532명 중 5,804명이 다시 S병원 내원하여 안과정밀 검진을 받았으며 이들 중 일부는 백내장 질환 진단을 받았다. 5,804명의 검진자는 적게는 1회에서 최대 12회까지 건강검진을 받았는데, 본 연구에서는 가장 최근의 검진 결과를 가지고 분석하였다.

백내장 발생 판별 모형을 추정하기 위한 반응 변수는 '백내장 발생자' 와 '백내장 비발생자'로 나뉘었으며 5,804명 중 백내장 발생자는 974명이고 비발생자는 4,830명으로 써 아래의 표 2와 같다.

표 2. 백내장 발생의 분포

| | 빈도 | 퍼센트 | 총합 |
|---|---|---|---|
| 백내장 발생 | 974 | 16.78% | 5,804 |
| 백내장 비발생 | 4,830 | 83.22% | |

모형 추정을 위해 사용된 독립 변수는 건강 검진 항목에서 기초정보, 신체 계측, 혈액검사, 전해질 검사, 뇨 검사, 간기능 검사, 지질 검사 변수들을 사용하였다. 총 26개의 독립변수가 분석에 사용 되었으며 독립변수들과 그것들의 기능에 대한 내용은 아래의 표3과 같다

표 3. 백내장 발생의 독립변수

| 검사 | 독립변수 | | 기능 |
|---|---|---|---|
| 기초 정보 | 성별 | | |
| | 연령 | | |
| 신체 계측 | BMI | $\frac{Weight}{Height^2}$ | 높을수록 심장질환, 당뇨병, 고혈압 등 성인병에 걸릴 확률이 높아진다. |
| 혈액검사 | WBC | 백혈구 수치 | 수치가 높으면 염증을 나타내는 질환 |
| | Platelet | 혈소판 수치 | 감소할수록 간 기능에 염증을 일으킬 확률이 커진다. |
| 대사 및 전해질 검사 | Na | 나트륨 | 나트륨을 과다 섭치하면 고혈압을 유발 |
| | K | 칼륨 | 많으면 신장에 영향 |
| | Ca | 칼슘 | 부족하면 백내장 유발 가능 |
| | P | 인 | 단백질・핵산 ・뼈・ 이의 성분, 신경・근육의 성분 |
| | Co2 | 체내 이산화탄소량 | 많을수록 호흡기 질환 발생 위험 |
| | Creatinine | 신장에 쌓인 노폐물 수치 | 높으면 신장에 이상 |
| 뇨 검사 | Glucose | 혈당 | 당뇨검사로 당뇨병이 있을 경우 높게 나타남 |
| | BUN | 혈중 요소질소 | 신장검사, 높으면 신부전증, 신우염 |
| | Ph | 소변 속의 산성도 측정 | 높으면 신장 비뇨기질환 |
| | Uro-bilirubin | 오줌 속에 포함된 bilirubin | 높으면 신장에 이상증상 |
| | Uric acid | 요산 | 혈중농도가 높으면 관절의 염증을 유발 |
| 간 기능 검사 | Protein | 간의 단백질 | |
| | Albumin | 알부민 | 저하 될수록 간 기능에 이상 |
| | Bilirubin | 담즙속의 적황색 색소 | 황달을 나타내는 수치. 이상 있을 시에 증가함 |
| | AST | 간효소 검사 | 증가할수록 간손상이 심함 |
| | ALT | 간효소 검사 | 간에만 존재하는 효소로 증가할수록 간손상이 심함 |
| | Alkaline phosphatase | 간세포 혈청 효소 | 간암,간경변,골질화,용혈성 항달이 있는 경우 증가 |
| | γ -GTP | 간세포 혈청 효소 | 간염,간경화 폐쇄성 활당 이상시 증가 |
| 혈청지질 | Cholesterol | 콜레스테롤 | 각종 심장질환과 노졸중 유발 |
| | Triglycerides | 중성지질 | 높을 경우 지방간 유발 |
| | HDL | 고농축 지지단백질 | 낮을수록 위험함. |

26개 독립변수의 일변량 분석의 결과는 표 4와 같다.

표 4. 독립변수의 일변량 분석 결과

| 변수 | 비발생 | | | 발생 | | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | ± | Std. Dev | Mean | ± | Std. Dev | p-value | |
| age | 50.29 | ± | 13.28 | 60.42 | ± | 10.71 | < 0.0001 | ** |
| sex (M/F) | 1759 | / | 2126 | 336 | / | 370 | 0.2559 | |
| BMI | 23.93 | ± | 3.78 | 24.19 | ± | 3.04 | 0.0467 | * |
| WBC | 6.37 | ± | 2.50 | 6.64 | ± | 1.91 | 0.0009 | ** |
| platelet | 241.65 | ± | 59.14 | 237.30 | ± | 62.98 | 0.0886 | |
| Na | 141.56 | ± | 4.49 | 141.66 | ± | 2.27 | 0.3784 | |
| K | 4.19 | ± | 0.37 | 4.25 | ± | 0.39 | < 0.0001 | ** |
| Ca | 9.49 | ± | 0.55 | 9.44 | ± | 0.47 | 0.0107 | ** |
| P | 3.60 | ± | 0.52 | 3.57 | ± | 0.54 | 0.2640 | |
| Co2 | 25.59 | ± | 2.53 | 25.56 | ± | 2.53 | 0.7459 | |
| Creatinine | 0.95 | ± | 0.23 | 0.99 | ± | 0.35 | 0.0103 | * |
| Glucose | 98.00 | ± | 29.35 | 114.32 | ± | 46.29 | < 0.0001 | ** |
| BUN | 14.37 | ± | 4.06 | 15.68 | ± | 4.80 | < 0.0001 | ** |
| Ph | 5.69 | ± | 0.96 | 5.67 | ± | 0.99 | 0.5927 | |
| Uro-bilirubin | 0.13 | ± | 0.25 | 0.14 | ± | 0.20 | 0.1311 | |
| Uric acid | 4.87 | ± | 1.41 | 4.87 | ± | 1.34 | 0.9338 | |
| Protein | 7.37 | ± | 0.47 | 7.34 | ± | 0.46 | 0.0783 | |
| Albumin | 4.57 | ± | 0.33 | 4.49 | ± | 0.35 | < 0.0001 | ** |
| Bilirubin | 0.79 | ± | 0.37 | 0.77 | ± | 0.50 | 0.3468 | |
| AST | 22.64 | ± | 22.87 | 24.00 | ± | 39.16 | 0.3695 | |
| ALT | 24.33 | ± | 36.93 | 23.99 | ± | 39.66 | 0.8351 | |
| Alkaline phosphatase | 72.72 | ± | 27.76 | 79.58 | ± | 32.19 | < 0.0001 | ** |
| γ-GTP | 30.54 | ± | 38.43 | 34.00 | ± | 51.59 | 0.0896 | |
| Cholesterol | 197.48 | ± | 37.30 | 200.99 | ± | 38.69 | 0.0224 | * |
| Triglycerides | 145.10 | ± | 100.49 | 157.65 | ± | 111.67 | 0.0054 | ** |
| HDL | 51.64 | ± | 12.87 | 50.03 | ± | 12.08 | 0.0027 | ** |

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

일변량 분석 결과 Age, BMI, WBC, K, Ca, Creatinine, Glucose, Bun, Albumin, Alkaline phosphatase ,Cholesterol, Triglycerides ,HDL 등 13개의 변수가 백내장 발생에 유의한 영향을 미치고 있음을 알 수 있다.

### 4.3.2 결측 자료 처리

백내장 발생 예측 모형에 사용된 5,804명의 자료의 독립변수들의 결측치를 조사한 결과 26개의 변수에서 20.90 %의 결측치가 있었다. 이는 1994년부터 2005년까지 건감검진 자료이므로 검사 항목이 도중에 추가되었거나 삭제되었기 때문이다. 본 논문에서 결측치 처리를 하지 않고 분석에 제외하였다. 따라서 5,804명 중 4,591명의 자료가 모형 추정에 사용되었다.

### 4.3.3 비용-민감도 달성

모형 추정에 사용된 4,591명 중 발생자와 비발생자의 비율은 다음 표5와 같다.

표 5. 결측치를 제외한 백내장 발생의 분포

| | 빈도 | 퍼센트 | 총합 |
|---|---|---|---|
| 백내장 발생자 | 707 | 15.40% | 4,591 |
| 백내장 비발생자 | 3884 | 84.60% | |

백내장 비발생자의 수가 발생자의 수에 비해 5배 정도의 높은 비율을 차지한다. 이와 같은 자료의 불균형은 민감도 (sensitivity)를 떨어뜨리는 경향이 있다. 또한 단순히 오분류 확률(misclassification probability)을 최소화 하는 방법으로 분류를 시행한다면 오분류 비용을 무시한 기준이므로 분류를 결정 할 때 문제가 발생할 수 있다. 분류의 목적은 백내장에 걸릴 위험이 있는 위험자를 더 정확하게 분류해 내는 것에 있으므로 민감도를 높이는 작업이 필요하다. 따라서 최적의 분류기준은 오분류 확률과 더불어 오분류 비용(misclassification cost)을 고려하는 것이 바람직하다. 자료의 불균형 문제 해결과 오분류 비용 적용을 위해 Charles Elkan et al 2001에 의해 소개된 이론을 사용하기로 한다. 비용 행렬(cost matrix)은 아래 표 6과 같이 정의되며 행은 예측을, 열은 실제값을 말한다.

표 6. 비용행렬

| | | actual | |
|---|---|---|---|
| | | negative | positive |
| Predict | negative | $C(0,0) = c_{oo}$ | $C(0,1) = c_{01}$ |
| | positive | $C(1,0) = c_{10}$ | $C(1,1) = c_{11}$ |

$c_{10}$ 은 false positive의 비용을 의미하며 $c_{01}$은 false negative의 비용을 의미한다. 실제 백내장 발생자을 비발생자로 분류하는 비용이 실제 비발생자를 발생자로 분류하는 비용에 비해 더 크므로 백내장 발생 모형을 예측하기 위해 사용된 비용행렬은 아래와 같이 정의 할 수 있다. 실제 발생자을 비발생자로 분류하는 비용은 실제 실제 비발생자를 발생자로 분류하는 비용의 5배라고 정의하면 비용행렬은 다음의 표 7과 같이 나타낸다.

표 7. 백내장 발생의 비용 행렬

| | 백내장 비발생자 | 백내장 발생자 |
|---|---|---|
| 백내장 비발생자 예측 | 0 | 5 |
| 백내장 발생자 예측 | 1 | 0 |

표준 분류 알고리즘은 정확도를 최대로 높이는 분계점에서 분류하도록 만들어진다. 두개의 그룹을 분류하는 경우 분류자는 확률 분계점 0.5에 기초하여 분류한다. 하지만 오분류 비용을 고려하게 되면 확률 분계점 0.5와는 다른 값을 가지게 된다. 즉 $x$가 주어졌을 때 $P(j=1|x) \geq p^*$인 목표 분계 점 $p^*$를 가지는 분류자가 필요하게 되는데 가장 일반적인 방법은 분석용 자료(training set)의 positive 와 negative의 수를 변화시키는 것이다. 분석용 자료의 negative의 수는

$$\frac{p^*}{1-p^*} \times \frac{1-p_0}{p_0} \times N_{negative}$$

으로 정의한다. 여기서 $p_0$ 은 주어진 확률 분계점이고, $p^*$은 목표 확률 분계점

을 말한다.

두 그룹으로 분류하는 경우 class 1로 예측할 최적인 방법은 class 0으로 예측할 기대비용보다 class 1으로 예측할 기대비용이 더 작거나 같을 경우를 말한다. 즉, 다음과 같이

$$P(j=0|x)c_{10} + P(j=1|x)c_{11} \leq P(j=0|x)c_{00} + P(j=1|x)c_{01}$$

표현한다. 위 식에서 $P(j=1|x) = p$라고 하면 아래와 같이

$$(1-p)c_{10} + pc_{11} \leq (1-p)c_{00} + pc_{01}$$

표현할 수 있다. 최적의 결정을 내리기 위한 목표 분계점은 서로 같을 때이므로 다음과 같이

$$(1-p^*)c_{10} + pc_{11} \leq (1-p^*)c_{00} + p^* c_{01}$$

표현한다. 위 식을 $p^*$에 대해서 정리하면 아래와 같은

$$p^* = \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}}$$

과 같다. 백내장 발생자의 자료의 경우 $p_0$이 0.5이고, $c_{00} = c_{11} = 0$이므로 $\frac{p^*}{1-p^*} = \frac{c_{10}}{c_{01}}$이 된다. 따라서 분석용 자료의 negative 수는 $\frac{c_{10}}{c_{01}} \times N_{negative} = \frac{1}{5} \times 3884 \fallingdotseq 777$ 명이다. 위와 같이 분석용 자료의 negative 수를 변화시키는 방법을 이용하여 자료의 불균형 문제를 해소하였다.

### 4.3.4 백내장 발생 예측 모형

백내장 발생 예측 모형은 '백내장 발생자'와 '백내장 비발생자' 두 그룹으로 구분하여 사건이 발생한 군은 '백내장 질환자'로 가정한다. 즉, 두 모집단을 가진 판별모형을 가정한다. 위 가정 하에 로지스틱 회귀분석의 혼합모형, 로지스틱 회귀모형, LDA, QDA를 이용하여 백내장 발생 예측모형을 각각 추정하고 예측확률을 구하여 이들 모형의 우수성을 비교할 것이다. 백내장 발생에 대한 예측 확률값은 confusion matrix로 작성할 수 있고 그것은 다음의 표 8와 같다.

표 8. Confusion Matrix

| | | Predict | |
|---|---|---|---|
| | | 백내장 비발생자 | 백내장 발생자 |
| Actual | 백내장 비발생자 | Specificity | False positive |
| | 백내장 발생자 | False negative | Sensitivity |

실제 자료에서 '백내장 발생자'를 예측모형에서도 백내장 발생자로 판별하는 확률을 민감도라고 하고, '백내장 비발생자'를 예측모형으로 판별하는 경우를 특이도라고 한다. 모형의 정확도는 이 민감도와 특이도가 모두 높아야 높일 수 있다. 두 개의 종속변수를 갖는 모형은 두 집단의 크기가 다를 때 작은 쪽의 집단에 왜곡된 결과를 초래할 수 있으므로 두 집단에 존재하는 오분류 비용을 고려하여 민감도와 특이도를 높이는 적절한 수준을 찾아야 한다. 이는 분석자에 의해 결정되어지게 된다.

# 제 5 장 혼합 로지스틱 분포를 이용한 백내장 예측 모형

## 5.1 모형의 설정

S병원 건강검진센터에서 1994년 5월 30일 ~ 2005년 9월 30일 사이에 건강검진을 받은 후 다시 내원하여 안과정밀 검사를 받은 4,591명의 검진자를 대상으로 하였다. 본 논문에서는 R (http://www.r-project.org)을 이용하여 프로그래밍을 하였으며, EM의 초기값을 계산하기 위해 MCLUST 라이브러리를 이용하였다.

백내장 발생의 판별 분석을 하기 전에 가장 영향을 주는 위험 인자를 추정하였다. 다음은 일변량 분석 결과 유의한 변수 13개 중 연령를 제외한 12개의 임상자료를 이용하였다. 그리고 변수들은 측정 기준이 모두 다르므로 각 변수마다 표준화시키기 위하여 각각의 평균으로 뺀 후 표준편차로 나누어 로지스틱 회귀분석을 실시하였으며 그 결과는 표9와 같다.

표 9 . 백내장 발생의 로지스틱 회귀분석 결과

| Parameter | DF | Estimate | Std.err | Chi-square | P-value | |
|---|---|---|---|---|---|---|
| Intercept | 1 | -1.786 | 0.0434 | 1690.16 | < 0.0001 | |
| BMI | 1 | 0.00869 | 0.0408 | 0.0454 | 0.8313 | |
| WBC | 1 | 0.0279 | 0.0367 | 0.5754 | 0.4481 | |
| K | 1 | 0.1265 | 0.0441 | 8.2184 | 0.0041 | ** |
| Ca | 1 | -0.1218 | 0.0517 | 5.5554 | 0.0184 | * |
| Creatinine | 1 | -0.029 | 0.044 | 0.4339 | 0.5101 | |
| Glucose | 1 | 0.293 | 0.0356 | 67.578 | < 0.0001 | ** |
| Bun | 1 | 0.2047 | 0.0445 | 21.1198 | < 0.0001 | ** |
| Albumin | 1 | -0.1627 | 0.0454 | 12.8386 | 0.0003 | ** |
| Alkaline phosphatse | 1 | 0.1168 | 0.0438 | 7.1083 | 0.0077 | ** |
| Cholestetol | 1 | 0.0112 | 0.0466 | 0.0574 | 0.8106 | |
| HDL | 1 | -0.015 | 0.0503 | 0.0891 | 0.7653 | |
| Triglycerides | 1 | 0.0442 | 0.0461 | 0.9215 | 0.3371 | |

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

일변량 분석에서 유의한 변수를 로지스틱 회귀모형에 적용한 결과 , K, Ca, Glucose Bun, Albumin, Alkaline phosphatase이 유의한 영향을 미침을 알 수 있다. 6개의 유의한 변수를 이용하여 가장 백내장 발생에 영향을 미칠 수 있는 변수들의 조합을 고려하기 위해서 Furnival and Wilson et al (1974)이 제안한 방법을 사용하였다. 그것은 모형에 포함될 독립변수를 선택하기 위해서 독립변수의 개수에 따라 가능한 모든 조합을 적용시킨 후 독립변수의 개수에 따라 likelihood score( 또는 Chi-square)값이 큰 값을 갖는 모형을 각각 선택하는 방법이다. 변수의 개수를 하나씩 증가 시켰을 때 Chi-square값이 가장 높은 부분집단의 결과는 표 10과 같다.

표 10. 독립변수들의 부분집단과 Chi-square

| 독립변수의 개수 | Score Chi-square | 포함된 변수 |
|---|---|---|
| 1 | 146.3034 | Glucose |
| 2 | 180.7986 | Glucose , Bun |
| 3 | 209.4626 | Glucose , Bun , Albumin |
| 4 | 218.0328 | Glucose , Bun , Albumin , Alkaline phosphatase |
| 5 | 222.6720 | Glucose , Bun , Albumin , Alkaline phosphatase , K |
| 6 | 229.4083 | Glucose , Bun , Albumin , Alkaline phosphatase , K , Ca |

1개의 독립변수를 선택할 경우 Glucose가 가장 높은 값을 가지고, 2개를 선택할 경우 Glucose 와 Bun의 조합이 가장 큼을 알 수가 있다. 본 논문에서는 독립변수가 하나일 때와 6개일 때의 경우를 제외하고, 2개~ 5개일 때의 모형을 이용하였다. 아래의 표 11는 독립변수의 개수에 따른 모형을 나타낸 것이다.

표 11. 독립변수의 개수에 따른 모형

| Model | Model Covariate |
|---|---|
| M1 | Glucose , Bun |
| M2 | Glucose , Bun , Albumin |
| M3 | Glucose , Bun , Albumin , Alkaline phosphatase |
| M4 | Glucose , Bun , Albumin , Alkaline phosphatase , K |

그리고 자료의 불균형 문제를 해소하기 위해 Cost 값을 1배, 3배, 5배로, 혼합 로지스틱 모형과 기존의 판별기법인 선형판별분석, 이차판별분석과의 모형평가를 위해 4,591명의 백내장 자료를 3:1 비율로 분석용 자료와 검증용 자료를 나누어 적용하였다.

## 5.2 혼합 로지스틱 회귀모형 적용

### 5.2.1 혼합 로지스틱 회귀모형의 모수 추정

독립변수의 개수에 따라 EM 알고리즘을 통해 성분에서의 혼합 비율과 $\beta_{ij}$ 을 추정하였다. 집단의 수는 1과 2로 미리 정해 놓고 그것에 따라 EM 알고리즘의 추정 방법에 따라서 수렴이 될 때까지 반복시켰다. 이때의 허용한계는 $10^{-6}$ 으로, 최대 반복수는 2000회로 설정하였다. 독립변수의 개수, Cost 값에 따라 적용한 혼합 로지스틱 회귀모형의 적합 결과는 아래의 표 12, 13, 14 이다.

표 12. 혼합 로지스틱 회귀모형의 적합 결과 ($g = 1 \sim 2$) : Cost 1배

| Model | $g$ | $\hat{\pi}$ | $\hat{\beta}$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Intercept | Glucose | Bun | Albumin | Alk,p | K |
| M1 | 1 | 1 | -1.788 | 0.329 | 0.199 | | | |
| | 2 | 0.385 | -1.143 | 1.989 | -0.155 | | | |
| | | 0.615 | -2.535 | -0.797 | 0.638 | | | |
| M2 | 1 | 1 | -1.797 | 0.325 | 0.188 | -0.230 | | |
| | 2 | 0.386 | -1.315 | 1.884 | -0.263 | 0.085 | | |
| | | 0.614 | -2.402 | -0.502 | 0.642 | -0.583 | | |
| M3 | 1 | 1 | -1.798 | 0.313 | 0.185 | -0.222 | 0.065 | |
| | 2 | 0.390 | -0.951 | 1.923 | -0.132 | -0.068 | 0.195 | |
| | | 0.610 | -3.043 | -1.299 | 0.730 | -0.538 | 0.064 | |
| M4 | 1 | 1 | -1.799 | 0.310 | 0.176 | -0.219 | 0.065 | 0.056 |
| | 2 | 0.387 | -0.926 | 2.048 | -0.126 | -0.082 | 0.189 | -0.036 |
| | | 0.613 | -3.143 | -1.600 | 0.711 | -0.507 | 0.084 | 0.250 |

표 13. 혼합 로지스틱 회귀모형의 적합 결과 ($g = 1 \sim 2$) : Cost 3배

| Model | $g$ | $\hat{\pi}$ | $\hat{\beta}$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Intercept | Glucose | Bun | Albumin | Alk,p | K |
| M1 | 1 | 1 | -0.632 | 0.398 | 0.225 | | | |
| | 2 | 0.355 | -2.967 | -3.777 | 1.544 | | | |
| | | 0.645 | -0.052 | 0.253 | -1.001 | | | |
| M2 | 1 | 1 | -0.647 | 0.340 | 0.206 | -0.228 | | |
| | 2 | 0.348 | 1.359 | 2.539 | 0.938 | -0.540 | | |
| | | 0.652 | -1.544 | 0.365 | 0.074 | -0.242 | | |
| M3 | 1 | 1 | -0.657 | 0.374 | 0.206 | -0.215 | 0.115 | |
| | 2 | 0.312 | -1.168 | 0.852 | 0.872 | -1.174 | 0.256 | |
| | | 0.688 | -0.509 | 1.184 | 0.038 | 0.048 | 0.053 | |
| M4 | 1 | 1 | -0.658 | 0.372 | 0.198 | -0.215 | 0.116 | 0.046 |
| | 2 | 0.287 | -1.893 | -1.011 | 1.118 | -0.795 | 1.007 | 1.5651 |
| | | 0.713 | -0.428 | 1.082 | 0.068 | -0.140 | -0.067 | -0.275 |

표 14. 혼합 로지스틱 회귀모형의 적합 결과 ( $g = 1 \sim 2$) : Cost 5배

| Model | $g$ | $\hat{\pi}$ | $\hat{\beta}$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Intercept | Glucose | Bun | Albumin | Alk,p | K |
| M1 | 1 | 1 | -0.249 | 0.361 | 0.245 | | | |
| | 2 | 0.224 | 0.012 | -1.310 | 2.443 | | | |
| | | 0.776 | -0.272 | 1.036 | -0.052 | | | |
| M2 | 1 | 1 | -0.264 | 0.368 | 0.240 | -0.255 | | |
| | 2 | 0.326 | 3.361 | -0.971 | 2.877 | 3.518 | | |
| | | 0.674 | -1.332 | 1.226 | -0.040 | -1.590 | | |
| M3 | 1 | 1 | -0.265 | 0.351 | 0.240 | -0.250 | 0.033 | |
| | 2 | 0.307 | 3.539 | -1.079 | 3.121 | 3.647 | 0.110 | |
| | | 0.693 | -1.180 | 1.081 | -0.024 | -1.363 | 0.278 | |
| M4 | 1 | 1 | -0.265 | 0.348 | 0.226 | -0.245 | 0.035 | 0.084 |
| | 2 | 0.299 | 4.511 | -1.259 | 4.682 | 4.063 | -0.107 | -1.398 |
| | | 0.701 | -1.119 | 0.987 | -0.806 | -1.236 | 0.320 | 0.379 |

Cost 1배일 경우, 집단이 두 개인 혼합 로지스틱 모형에서의 혼합비율은 각각 0.39, 0.61로 이루어져 있다. 0.39의 혼합비율을 가진 집단의 회귀계수 추정값은 Glucose(혈당수치), Albumin (간 기능 수치)이 양의 값, 0.61의 혼합비율을 가진 집단의 회귀계수 추정값은 BUN(혈중 요소 질소), K이 양의 값으로 나타났다. 그리고 Alkaline phosphatase은 각각의 집단에서 양의 값으로 나타났다. 이것은 혈

당수치나 간 기능 수치가 클수록 백내장 발생에 영향을 주고, 0.61의 혼합 비율을 가진 집단은 혈중 요소 질소 와 칼륨, 즉 대사 및 전해질의 양이 클수록 백내장 발생에 영향을 주는 것을 알 수 있다.

Cost가 3배, 5배일 경우에도 각 집단에 따라 영향을 주는 변수가 다름을 살펴 볼 수 있다. 따라서 집단이 1개일 경우의 일반적인 로지스틱 회귀모형으로부터 추정된 회귀계수 보다 집단이 2개인 혼합 로지스틱 모형이 백내장 발생에 영향을 줄 수 있는 위험요인의 복합적인 영향에 대해 설명할 수 있음을 알 수 있다.

### 5.2.2 집단의 수 결정

EM 알고리즘을 통해 각 집단 또는 성분의 수를 따라 알려지지 않은 모수들을 추정한 후에 AIC의 기준에 맞추어 과연 몇 개의 혼합 로지스틱 모형으로 이루어져 있는지를 선택할 수 있다. 그 판단의 기준은 BIC 또는 AIC의 값이 최초에 최소가 되는 값을 선택한다.

표 15는 각 집단의 혼합 로지스틱 회귀모형을 통해 구한 BIC와 AIC의 추정치를 나타낸 것이다. 표 15를 살펴보면, 로그 우도함수 값은 집단이 1개일 때 보다 2개일 경우 커지며, 또한 독립변수의 개수가 증가할수록 커짐을 볼 수 있다. AIC의 값은 집단이 1개일 때 보다는 2개일 때 감소함을 알 수 있다. BIC의 값은 집단이 1개 또는 2개일 때 차이가 별로 나지 않음을 볼 수 있다.

따라서 AIC의 값을 기준으로 하면, 이 모형에서는 집단이 2개임을 알 수 있다. 즉, 혼합 로지스틱 모형의 분포가 2개의 집단으로 이루어져 있음을 알 수 있다. 2개의 집단에서 Cost 1배 일 경우에는 M2모형 ,Cost 3배일 경우에는 M3 모형, Cost가 5배 일 경우에는 M4모형이 가장 작은 AIC값을 가짐을 볼 수 있다. 이것은 Cost에 따라 백내장 질환자의 비율이 커지고 그것을 설명할 변수가 더 필요함을 알 수 있다.

표 15. 집단 ($g = 1, 2$) 의 변화에 따른 BIC, AIC 결과

I. M1, M2 모형

| Cost | g | M1 | | | M2 | | |
|---|---|---|---|---|---|---|---|
| | | BIC | AIC | logL | BIC | AIC | logL |
| 1 | 1 | 2833.17 | 2814.74 | -1404.37 | 2820.65 | 2796.08 | -1394.04 |
| | 2 | 2828.28 | 2785.27 | -1385.63 | 2820.48 | 2765.18 | -1373.59 |
| 3 | 1 | 1781.80 | 1766.06 | -880.03 | 1774.20 | 1753.21 | -872.60 |
| | 2 | 1789.90 | 1753.17 | -869.59 | 1794.28 | 1747.06 | -864.53 |
| 5 | 1 | 1488.83 | 1473.79 | -733.89 | 1480.67 | 1460.62 | -726.31 |
| | 2 | 1502.76 | 1467.66 | -726.83 | 1487.30 | 1442.18 | -712.09 |

II. M3, M4 모형

| Cost | g | M3 | | | M4 | | |
|---|---|---|---|---|---|---|---|
| | | BIC | AIC | logL | BIC | AIC | logL |
| 1 | 1 | 2826.56 | 2795.84 | -1392.92 | 2833.47 | 2796.61 | -1392.30 |
| | 2 | 2833.21 | 2765.62 | -1371.81 | 2846.35 | 2766.47 | -1370.24 |
| 3 | 1 | 1777.49 | 1751.25 | -870.63 | 1784.13 | 1752.65 | -870.33 |
| | 2 | 1804.90 | 1747.19 | -862.59 | 1818.02 | 1749.81 | -861.90 |
| 5 | 1 | 1487.27 | 1462.21 | -726.10 | 1492.60 | 1462.52 | -725.26 |
| | 2 | 1496.61 | 1441.45 | -709.73 | 1504.99 | 1439.81 | -706.90 |

### 5.2.3 로지스틱 회귀모형과의 적합도 비교

3장에서 모형을 평가하기 위해 설정된 모형이 자료를 얼마나 잘 적합하고 있는지 여부를 살펴보아야 한다고 하였다. 집단이 1개 일 때는 로지스틱 회귀모형이므로, 이것에 비해 2개의 집단으로 이루어진 혼합 로지스틱회귀모형이 실제자료에 얼마나 더 가까운지를 살펴보아야 한다.

이 때 판단의 기준은 데비언스를 이용하고 이 값이 클수록 적합도가 떨어지므로 작은 값을 가지는 모형이 더 예측을 잘 한다고 판단할 수 있다. 분석용 자료와 검증용 자료에서 각 집단에서의 혼합 로지스틱 회귀모형의 데비언스 값은 표 16과 같다.

표 16을 보면 분석용 자료에서는 집단이 2개인 혼합 로지스틱이 데비언스 값이 모두 작았다. 검증용 자료에서는 Cost 1인 경우를 제외 하고 혼합 로지스틱 회귀모형일 때 데비언스 값이 로지스틱 회귀모형보다 작음을 알 수 있다. Cost 1인 경우 백내장 질환자의 비율이 15%정도로 불균형이 심하기 때문에 로지스틱 회귀모형이 혼합 로지스틱 모형보다 작은 데비언스 값을 가지는 것으로 고려되어지지만 변수가 5개인 M4모형일 경우에는 데비언스 값이 로지스틱 회귀모형보다 작아짐을 볼 수 있다. 추정된 데비언스 값을 보았을 때 로지스틱 회귀모형보다는 두개의 혼합 로지스틱 모형이 좀더 실제자료에 가까워지는 것을 볼 수 있다.

표 16. 집단 ($g = 1, 2$) 의 변화에 따른 데비언스

I. 분석용 자료

| Cost | g | M1 | M2 | M3 | M4 |
|---|---|---|---|---|---|
| 1 | 1 | 2808.47 | 2788.08 | 2785.843 | 2784.61 |
| | 2 | 2771.27 | 2747.18 | 2743.622 | 2740.47 |
| 3 | 1 | 1760.06 | 1745.21 | 1741.253 | 1740.65 |
| | 2 | 1739.17 | 1729.06 | 1725.185 | 1723.81 |
| 5 | 1 | 1467.79 | 1452.62 | 1452.205 | 1450.52 |
| | 2 | 1453.662 | 1424.18 | 1419.454 | 1413.81 |

II. 검증용 자료

| Cost | g | M1 | M2 | M3 | M4 |
|---|---|---|---|---|---|
| 1 | 1 | 929.892 | 985.977 | 981.578 | 978.459 |
| | 2 | 985.307 | 988.754 | 984.261 | 978.312 |
| 3 | 1 | 617.603 | 616.649 | 617.065 | 616.153 |
| | 2 | 615.896 | 619.656 | 615.175 | 612.12 |
| 5 | 1 | 503.914 | 499.973 | 499.011 | 496.998 |
| | 2 | 497.814 | 497.532 | 495.076 | 491.175 |

## 5.3 기존 분석과의 비교

2장에서 제시하고 있는 기존에 많이 사용되어지고 있는 선형판별분석와 이차판별분석, 로지스틱 회귀분석을 본 논문에서 제시한 혼합 로지스틱 회귀모형을 모형의 평가기준항목인 민감도, 특이도, 정확도 등으로 각 모형을 평가해 보고자 한다.

마찬가지로 분석용 자료와 검증용 자료에서 독립변수의 개수에 따라 Cost값을 1배, 3배, 5배로 변화시켜가면서 모형을 추정하여 평가해 보았다. 그 결과는 표 17, 18과 같다.

표 17. 백내장 발생 예측 모형의 분석용 자료의 분류표

| Model | Method | cost | sensitivity | specificity | false positive | false negative | Accuracy | Mis classification |
|---|---|---|---|---|---|---|---|---|
| M1 | LDA | 1 | 5.81 | 98.74 | 55.22 | 14.40 | 84.81 | 15.19 |
| | | 3 | 17.09 | 94.49 | 35.77 | 33.70 | 66.10 | 33.90 |
| | | 5 | 29.41 | 84.39 | 38.52 | 41.47 | 59.17 | 40.83 |
| | QDA | 1 | 12.40 | 96.65 | 60.49 | 13.78 | 84.03 | 15.97 |
| | | 3 | 20.97 | 93.36 | 35.33 | 32.90 | 66.81 | 33.19 |
| | | 5 | 24.90 | 89.37 | 33.51 | 41.59 | 59.80 | 40.20 |
| | Mixture (g=1) | 1 | 2.13 | 99.25 | 66.67 | 14.81 | 84.69 | 15.31 |
| | | 3 | 17.48 | 94.38 | 35.71 | 33.62 | 66.17 | 33.83 |
| | | 5 | 30.20 | 83.72 | 38.89 | 41.40 | 59.17 | 40.83 |
| | Mixture (g=2) | 1 | 0.58 | 99.90 | 50.00 | 14.93 | 85.01 | 14.99 |
| | | 3 | 27.18 | 88.30 | 42.62 | 32.33 | 65.88 | 34.12 |
| | | 5 | 34.90 | 81.06 | 39.04 | 40.49 | 59.89 | 40.11 |
| M2 | LDA | 1 | 6.59 | 98.60 | 54.67 | 14.31 | 84.81 | 15.19 |
| | | 3 | 19.42 | 93.70 | 35.90 | 33.25 | 66.45 | 33.55 |
| | | 5 | 36.08 | 84.72 | 33.33 | 39.00 | 62.41 | 37.59 |
| | QDA | 1 | 13.76 | 96.28 | 60.56 | 13.64 | 83.91 | 16.09 |
| | | 3 | 22.91 | 92.91 | 34.81 | 32.46 | 67.24 | 32.76 |
| | | 5 | 18.59 | 88.87 | 31.60 | 54.27 | 49.20 | 50.80 |
| | Mixture (g=1) | 1 | 2.52 | 99.15 | 65.79 | 14.77 | 84.66 | 15.34 |
| | | 3 | 19.81 | 93.36 | 36.65 | 33.23 | 66.38 | 33.62 |
| | | 5 | 38.24 | 83.72 | 33.45 | 38.46 | 62.86 | 37.14 |
| | Mixture (g=2) | 1 | 0.97 | 99.76 | 58.33 | 14.89 | 84.95 | 15.05 |
| | | 3 | 21.75 | 91.34 | 40.74 | 33.17 | 65.81 | 34.19 |
| | | 5 | 41.57 | 81.89 | 33.96 | 37.67 | 63.40 | 36.60 |
| M3 | LDA | 1 | 6.40 | 98.53 | 56.58 | 14.35 | 84.72 | 15.28 |
| | | 3 | 19.22 | 93.93 | 35.29 | 33.25 | 66.52 | 33.48 |
| | | 5 | 36.27 | 84.88 | 32.97 | 38.88 | 62.59 | 37.41 |
| | QDA | 1 | 13.57 | 96.28 | 60.89 | 13.66 | 83.88 | 16.12 |
| | | 3 | 23.69 | 91.11 | 39.30 | 32.67 | 66.38 | 33.62 |
| | | 5 | 31.76 | 87.71 | 31.36 | 39.73 | 62.05 | 37.95 |
| | Mixture (g=1) | 1 | 2.33 | 99.11 | 68.42 | 14.80 | 84.61 | 15.39 |
| | | 3 | 19.61 | 93.25 | 37.27 | 33.31 | 66.24 | 33.76 |
| | | 5 | 38.24 | 83.89 | 33.22 | 38.41 | 62.95 | 37.05 |
| | Mixture (g=2) | 1 | 0.97 | 99.80 | 54.55 | 14.89 | 84.98 | 15.02 |
| | | 3 | 24.66 | 90.10 | 40.93 | 32.63 | 66.10 | 33.90 |
| | | 5 | 41.37 | 82.23 | 33.65 | 37.66 | 63.49 | 36.51 |
| M4 | LDA | 1 | 6.40 | 98.53 | 56.58 | 14.35 | 84.72 | 15.28 |
| | | 3 | 23.69 | 90.89 | 39.90 | 32.72 | 66.24 | 33.76 |
| | | 5 | 35.29 | 84.22 | 34.55 | 39.43 | 61.78 | 38.22 |
| | QDA | 1 | 14.15 | 96.28 | 59.89 | 13.58 | 83.97 | 16.03 |
| | | 3 | 23.69 | 90.89 | 39.90 | 32.72 | 66.24 | 33.76 |
| | | 5 | 30.78 | 87.04 | 33.19 | 40.25 | 61.24 | 38.76 |
| | Mixture (g=1) | 1 | 2.52 | 99.18 | 64.86 | 14.77 | 84.69 | 15.31 |
| | | 3 | 19.81 | 93.14 | 37.42 | 33.28 | 66.24 | 33.76 |
| | | 5 | 39.02 | 82.72 | 34.32 | 38.44 | 62.68 | 37.32 |
| | Mixture (g=2) | 1 | 0.97 | 99.69 | 64.29 | 14.90 | 84.90 | 15.10 |
| | | 3 | 25.63 | 89.65 | 41.07 | 32.46 | 66.17 | 33.83 |
| | | 5 | 43.53 | 77.41 | 35.47 | 40.79 | 60.95 | 39.05 |

표 18. 백내장 발생 예측 모형의 검증용 자료의 분류표

| Model | Method | cost | sensitivity | specificity | false positive | false negative | Accuracy | Mis classification |
|---|---|---|---|---|---|---|---|---|
| M1 | LDA | 1 | 6.28 | 98.85 | 47.83 | 15.91 | 83.45 | 16.55 |
| | | 3 | 14.58 | 96.38 | 26.32 | 38.14 | 62.82 | 37.18 |
| | | 5 | 26.90 | 86.78 | 30.26 | 48.81 | 54.99 | 45.01 |
| | QDA | 1 | 13.61 | 95.82 | 60.61 | 15.25 | 82.14 | 17.86 |
| | | 3 | 16.15 | 95.29 | 29.55 | 37.97 | 62.82 | 37.18 |
| | | 5 | 22.34 | 90.23 | 27.87 | 49.35 | 54.18 | 45.82 |
| | Mixture (g=1) | 1 | 2.62 | 99.48 | 50.00 | 16.34 | 83.36 | 16.64 |
| | | 3 | 15.63 | 96.38 | 25.00 | 37.85 | 63.25 | 36.75 |
| | | 5 | 30.96 | 86.21 | 28.24 | 47.55 | 56.87 | 43.13 |
| | Mixture (g=2) | 1 | 0.52 | 99.90 | 50.00 | 16.58 | 83.36 | 16.64 |
| | | 3 | 23.44 | 90.58 | 36.62 | 37.03 | 63.03 | 36.97 |
| | | 5 | 36.04 | 83.91 | 28.28 | 46.32 | 58.49 | 41.51 |
| M2 | LDA | 1 | 8.38 | 98.43 | 48.39 | 15.67 | 83.45 | 16.55 |
| | | 3 | 14.58 | 95.29 | 31.71 | 38.41 | 62.18 | 37.82 |
| | | 5 | 32.49 | 86.78 | 26.44 | 46.83 | 57.95 | 42.05 |
| | QDA | 1 | 14.14 | 95.82 | 59.70 | 15.17 | 82.23 | 17.77 |
| | | 3 | 18.75 | 94.93 | 28.00 | 37.32 | 63.68 | 36.32 |
| | | 5 | 24.37 | 91.38 | 23.81 | 48.38 | 55.80 | 44.20 |
| | Mixture (g=1) | 1 | 13.61 | 95.82 | 60.61 | 15.25 | 82.14 | 17.86 |
| | | 3 | 16.67 | 95.29 | 28.89 | 37.83 | 63.03 | 36.97 |
| | | 5 | 32.26 | 85.06 | 27.08 | 49.83 | 55.75 | 44.25 |
| | Mixture (g=2) | 1 | 2.62 | 99.27 | 58.33 | 16.37 | 83.19 | 16.81 |
| | | 3 | 20.31 | 94.20 | 29.09 | 37.05 | 63.89 | 36.11 |
| | | 5 | 35.03 | 83.91 | 28.87 | 46.72 | 57.95 | 42.05 |
| M3 | LDA | 1 | 7.85 | 98.64 | 46.43 | 15.71 | 83.54 | 16.46 |
| | | 3 | 15.10 | 95.29 | 30.95 | 38.26 | 62.39 | 37.61 |
| | | 5 | 32.99 | 86.78 | 26.14 | 46.64 | 58.22 | 41.78 |
| | QDA | 1 | 14.14 | 95.51 | 61.43 | 15.21 | 81.97 | 18.03 |
| | | 3 | 17.19 | 94.20 | 32.65 | 37.95 | 62.61 | 37.39 |
| | | 5 | 25.38 | 90.23 | 25.37 | 48.36 | 55.80 | 44.20 |
| | Mixture (g=1) | 1 | 3.66 | 99.27 | 50.00 | 16.23 | 83.36 | 16.64 |
| | | 3 | 15.63 | 94.93 | 31.82 | 38.21 | 62.39 | 37.61 |
| | | 5 | 32.72 | 84.48 | 27.55 | 49.83 | 55.75 | 44.25 |
| | Mixture (g=2) | 1 | 1.05 | 99.37 | 75.00 | 16.58 | 83.01 | 16.99 |
| | | 3 | 21.88 | 93.48 | 30.00 | 36.76 | 64.10 | 35.90 |
| | | 5 | 36.04 | 81.03 | 31.73 | 47.19 | 57.14 | 42.86 |
| M4 | LDA | 1 | 7.85 | 98.43 | 50.00 | 15.74 | 83.36 | 16.64 |
| | | 3 | 17.71 | 93.48 | 34.62 | 37.98 | 62.39 | 37.61 |
| | | 5 | 34.01 | 86.21 | 26.37 | 46.43 | 58.49 | 41.51 |
| | QDA | 1 | 15.18 | 95.72 | 58.57 | 15.03 | 82.32 | 17.68 |
| | | 3 | 17.71 | 93.48 | 34.62 | 37.98 | 62.39 | 37.61 |
| | | 5 | 27.41 | 90.80 | 22.86 | 47.51 | 57.14 | 42.86 |
| | Mixture (g=1) | 1 | 3.66 | 99.48 | 41.67 | 16.20 | 83.54 | 16.46 |
| | | 3 | 16.15 | 95.29 | 29.55 | 37.97 | 62.82 | 37.18 |
| | | 5 | 36.04 | 84.48 | 27.55 | 46.15 | 58.76 | 41.24 |
| | Mixture (g=2) | 1 | 1.05 | 99.48 | 71.43 | 16.56 | 83.10 | 16.90 |
| | | 3 | 21.88 | 92.03 | 34.38 | 37.13 | 63.25 | 36.75 |
| | | 5 | 39.09 | 81.03 | 30.00 | 45.98 | 58.76 | 41.24 |

분석결과를 살펴보면, 분석용 자료와 검증용 자료 모두 Cost값이 높을수록 민감도는 높아지고 특이도와 정확도는 낮아짐을 볼 수 있다. 그러나 상대적으로 다른 모형에 비해 혼합 로지스틱 회귀모형이 민감도가 더 높게 나타나고 있으며, 특이도와 정확도 또한 높게 나타나고 있음을 알 수 있다. 이때, Cost 값이 3배일 때 민감도, 특이도, 정확도가 적정한 수준을 이루고 있다. 다음의 그림 1은 검증용 자료에서 Cost값이 3배일 때 독립변수에 따른 민감도, 특이도, 정확도 그림이다.

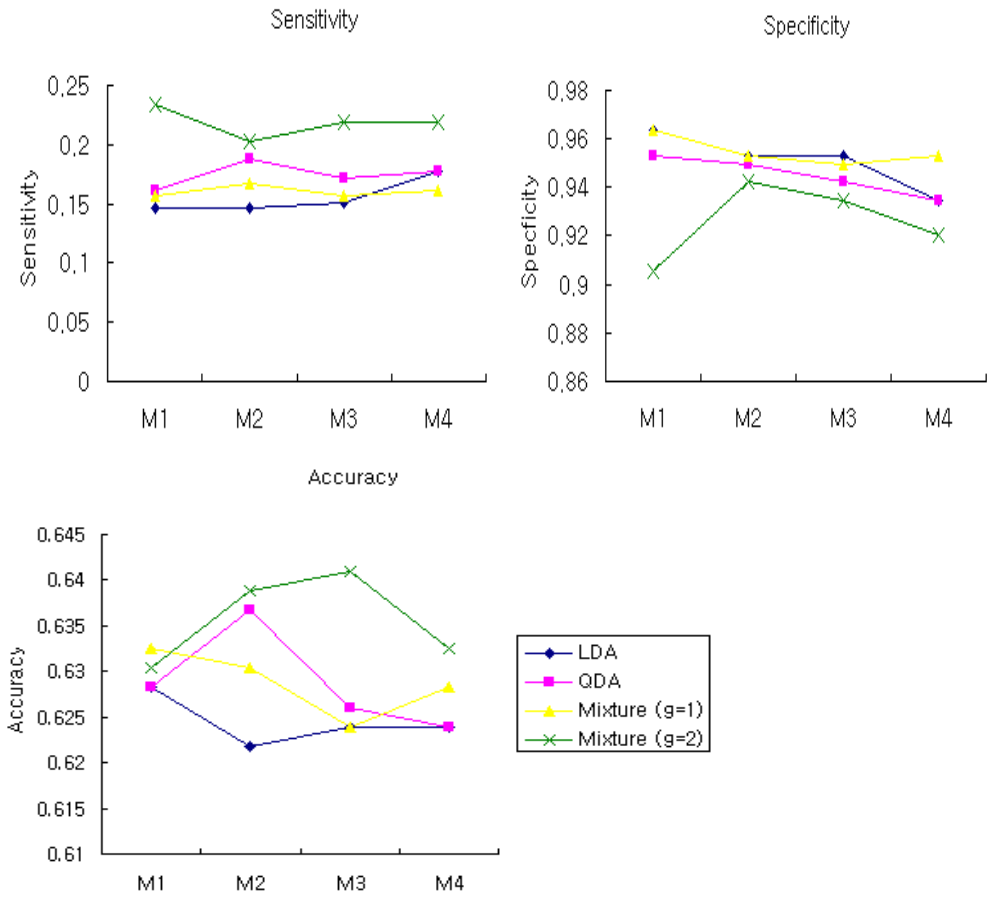


그림 1. 검증용 자료에서 Cost 3 일 때의 민감도, 특이도, 정확도

위 그림 1을 보면 최대 5개의 독립변수를 이용하는 것보다 독립변수가 4개인 M3모형일 때 민감도와 정확도가 높음을 알 수 있다. 독립변수의 개수가 많을수록 민감도와 특이도, 정확도를 높여주는 것은 아님을 알 수 있다. 따라서 Cost 값이 3배이고 독립변수가 4개인 M3모형에 대해 살펴보았는데, 그 결과는 표 19와 같다.

표 19. Cost 3배, 4개의 독립변수일 때의 모형의 분류표

| Set | Method | sensitivity | specificity | false positive | false negative | Accuracy | Mis classification |
|---|---|---|---|---|---|---|---|
| Training | LDA | 19.22 | 93.93 | 35.29 | 33.25 | 66.52 | 33.48 |
| | QDA | 23.69 | 91.11 | 39.30 | 32.67 | 66.38 | 33.62 |
| | Logistic | 19.61 | 93.25 | 37.27 | 33.31 | 66.24 | 33.76 |
| | Mixture | 24.66 | 90.10 | 40.93 | 32.63 | 66.10 | 33.90 |
| Test | LDA | 15.10 | 95.29 | 30.95 | 38.26 | 62.39 | 37.61 |
| | QDA | 17.19 | 94.20 | 32.65 | 37.95 | 62.61 | 37.39 |
| | Logistic | 15.63 | 94.93 | 31.82 | 38.21 | 62.39 | 37.61 |
| | Mixture | 21.88 | 93.48 | 30.00 | 36.76 | 64.10 | 35.90 |

표 20의 결과를 보면 기존의 알고리즘에 비해 혼합 로지스틱 모형이 오분류율이 낮음을 알 수 있다. 혼합로지스틱 모형이 21.88%로 다른 알고리즘에 비해 높게 나타났으며 정확도는 64.10%로 나타났다. 전체적으로 민감도와 정확도가 높지 않은 것은 안과정밀검사를 통해 얻어지는 자료가 아닌 검진자료를 통해 얻어지는 임상적 자료만을 대상으로 분석을 하였기 때문에 이런 결과가 나타난 것으로 고려되어진다.

# 제 6 장 토의 및 결론

건강검진은 만성질환을 일으킬 수 있는 위험요소나 신체변화를 조기발견과 현재의 건강상태를 알아보기 위한 목적으로 실시하고 있다. 본 논문에서는 통계적 자료 분석기법을 통해 질병에 대한 조기진단이라는 측면에서 백내장 발생에 관한 연구를 하였다. 백내장은 신체정보, 간수치, 대사 및 전해질 수치 및 여러 가지 복합적인 위험요인에 의해 발생되는 질병이므로 그 특성에 따라 여러 개의 하위집단으로 이루어져 있다고 볼 수 있다. 혼합 로지스틱 회귀모형은 여러 개의 집단으로 이루어진 자료에서 독립변수들이 이분형 종속변수에 영향을 미칠 때 고려할 수 있는 통계적 기법에 해당된다.

본 논문에서는 5,804명의 검진자 중 결측치를 제외한 4,591명을 대상으로 독립변수의 개수에 따라 집단이 1~2개인 혼합 로지스틱 회귀모형에 적용하였다. 모수추정에서 최대우도 추정량을 구하기 위하여 EM 알고리즘을 사용하였다. 추정된 모수를 이용하여 집단의 수는 AIC 값으로 결정하였고, 실제 자료에 어느 정도 가깝게 예측하는지를 데비언스 값으로 살펴보았다. 그 결과  하나의 로지스틱 회귀모형 보다는 집단이 두 개인 혼합 로지스틱 회귀모형을 적용하였을 때, 위험요인의 복합적인 영향에 대해 더 잘 설명할 수 있고  실제 백내장 발생에 대한 예측력이 높다는 것을 알 수 있었다. 또한, 선형판별분석과 이차판별분석을  통해 백내장 발생을 살펴 본 결과,  4개의 독립변수(Glucose , Bun , Albumin , Alkaline phosphatase)로 이루어진 조합에서 혼합로지스틱 회귀모형이 정확도  64.10%, 민감도 21.88% 으로 다른 기법에 비해 예측력이 높음을 확인하였다.

또한, 네 개의 독립변수는 이미 알려진 위험요인으로 알려진 혈당수치 이외에 간과 관련된 항목의 수치, 대사 및 전해질 수치, 신장 기능 수치 등 신체와 관련된 항목이 위험요인임을 알 수 있었다. 그 결과, 예측모형을 통해 의사의 진단과 안과 정밀 검사 없이 건강검진 자료만을 통해서 백내장 질환 유·무에 관한 정보를 64%정도 예측 할 수 있음을 알 수 있다.

본 논문을 토대로 건강검진 자료를 통해 백내장 질환에 대한 조기 진단을 할

수 있을 것으로 보이는데, 이를 바탕으로 모형을 실제 진단에 적용하는 연구가 이루어져야 할 것이다. 여기서 다루지 못한 독립변수들의 집단이 3개 이상인 경우 또한  살펴 볼 필요가 있다. 또한 결측치가 포함 되었을 경우 이것을 모형에서 제외하지 않고 백내장 발생을 예측하는데 포함시켜 분석하는 방법으로의 확장이 필요하며, 검진자의 반복적인 건감검진에 따른 백내장 발생에 관한 예측 방법에 관한 연구도 진행될 필요가 있다고 판단한다.

# 참 고 문 헌


김동해. 백내장의 발생 위험요인에 대한 환자 대조군 연구. 서울대 대학원 석사학위논문, 1999.

김재호. 백내장 백과사전. 일조각, 2002.

성웅현. 응용 로지스틱 회귀분석. 탐진, 2001.

이상욱. 백내장・녹재장. 민중서관, 1999.

최진석,김신자,장병원,신경환. 한국실명예방재단의 2004년도 노인 안 검진과 개안수술사업 결과. 대한안과학회지, 2005: 46(1); 63-70.

한은정. 건강검진 자료에서 Random Forest를 이용한 백내장 발생 위험군 예측모형. 연세대 대학원 석사학위 논문, 2004.

Chen,K.,Xu,L.,Chi,H. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*. 1999: 12;1229-1252.

Elkan,C. The foundations of cost-sensitive learning. *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (IJCAI'01), August 2001: 973-978.

Follmann,D.A.,Lambert,D. Generalizing Logistic Regression by Nonparametric Mixing. *Journal of American Statistical Association*, 1989: 84(405); 295-300.

Follmann,D.A. Identifiability of finite mixtures of logistic regression model. *Journal of Statistical Planning and Inference*,1991: 27 ; 375-381.

Fraley,C.,Raftery,A.E. Model-based Clustering, Discriminant Analysis, and Density Estimation. *Journal of American Statistical Association*. 2002: 97; 611-631.

Furnival,G.M.,Wilson,R.W. Regressions by Leaps and Bounds.*Technometrics,* 1974: 16; 499-511.

HastieT.,Tibshirani,R.*J.* Discriminant Analysis by Gaussian Mixture. *Journal of Royal Statistical Society Series* , 1996: 58 (1);155-176.

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

Hosmer,D.W.,Lemeshow.S. Aplied Logistic Regression. *Wiley*, 2002.

Huberty, Carl *J.* Applied Discriminant Analysis. *Wiley* , 1994.

*J*effries,N.O. Logistic Mixtures of Generalized Linear Model Times Series. PH.D Thesis, University of Maryland ,1998.

*J*ordan,M.I., *J*acobs,R.A. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*. 1994: 6(2);181-214.

McCullagh,P.,Nelder,*J*.A. Generalized Linear Models. *Chapman & Hall/CRC*, 1997.

McLachlan,G.*J.*, Krishnan, T. The EM algorithm and Extensions, *Wiley*,1997.

McLachlan,G.*J.*, Peel, D. Finite Mixture Models. *Wiley* ,2002

O'Neill, T.*J.* Mixture or logistic regression estimation for discrimination, *Statistical & Probability Letters* , 1994:20;139-142.

Titasias,M.K.,Likas,A. Mixture of Experts Classification Using a Hierarchical Mixture model. *Neural Computation*, 2002: 14; 2221-2244.

# ABSTRACT

## A Study of Prediction model for the Development of Cataract Using Logistic Mixtures

**Kim, So Yeon**
**Dept. of Biostatistics and Computing**
**The Graduate School**
**Yonsei University**

Cataracts are layers over a person's eyes that prevent them from seeing properly. However, development of cataracts can be reduced through early diagnosis from screening test. In this study, we predict the development of cataract based on screening data accumulated from 1994 to 2005, and identify risk factor related with cataract. We used logistic mixtures as prediction model and compared the performance of any other methods with ours. AIC was used as the selection criteria of component g and deviance as a measure of the fittability

When fitting mixture models of data, model with two-components had better performance than model with one, that is, the estimated coefficient vectors in two-components model explained the effect of complex risk factors well. The accuracy and sensitivity using logistic mixture were 64.10% and 21.88% and these were higher them those of any other methods. In these

results, we found that the logistic mixture model could improve the predictability of the development of cataract.

---

Key words : Screening test, cataract, Risk factor, Mixture logistics, EM algorithm, prediction model