데이터 마이닝을 위한 일반화 선형모형 구현 및 확장: 다구찌 품질개선실험 자료분석의 적용

이 영 조, 노 맹 석
서울대 통계학과
지 원 철
홍익대 산업공학과

1. 서론

최근 데이터 마이닝에 대한 관심은 데이터로부터 유용한 정보를 찾아내는 통계 기법에 대한 관심을 고조시켰다. 본 발표에서는 다구찌 품질개선실험자료분석 방법을 통해 통계학의 모형 및 분석법이 어떻게 변화하여 왔는지 또한 어떤 방향으로 나아가야 되는지에 대하여 논의하고, 그 구현을 위한 소프트웨어 개발 및 새로운 마이닝 방법을 제시하고자 한다.

2. 회귀분석과 자료변환

다구찌 품질실험 (1985, 1987) 으로 부터 얻어진 자료를 분석하는 기본적인 수단으로 회귀분석이 사용되는데 반응변수 y가 다음 세가지 조건을 만족한다고 가정한다.

(1) 정규성(Normality) : 반응변수 y가 정규분포를 따른다.

(2) 가법성(Additivity) : $\mu = E(y)$
$= \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k.$

(3) 등분산성 (Constant Variance) :
$$Var(y) = \sigma^2 I.$$

품질개선실험에서 불량품의 개수나 불량율등의 변수들은 포아송 분포나 이항 분포를 따르므로 정규성 가정을 만족하지 않으며, 또한 분산이 일정하지 않고 평균에 따라 변한다. 예를 들면 포아송 분포에서는 분산이 평균에 비례한다. 흔히 이런 자료들은 독립 변수들의 효과에 대한 가법모형(2)이 평균척도(μ-scale)에서 성립한다는 보장도 없다. 포아송 분포에서는 가법성이 평균의 로그척도에서 흔히 만족되므로 이 세 가지 가정의 일부 또는 모두를 만족하는 자료는 실제 매우 드물다. 그러므로 Box와 Cox(1964)는 자료가 위의 세 조건을 일부 또는 모두 만족시키지 않을 경우에는 이들을 충족시키도록 자료변환을 한 뒤 분석을 해야 한다고 주장하였다.

그러나 Hougaard(1982)가 밝혔듯이 이 세 가지 조건을 동시에 충족시킬 수 있는 자료변환은 존재하지 않는다. 간단한 예로 포아송 분포를 생각해 보자. 정규분포 가정을 위하여는 $y^{2/3}$. 등분산성 가정을 위하여는 $y^{1/2}$, 포아송자료에서 많이 발생하는 곱모형(Multiplicative Model)이 가법성을 만족하기 위하여는 $\log y$ 변환이 필요하다. 자료변환의 문제점은 이와 같이 세 조건을 충족시키기 위하여 각기 다른 세 가지 자료변환이 요구된다는 점이다.

3. 일반화 선형모형

일반화 선형모형은 회귀분석의 세가지 가정을 확장하여 좀더 일반적인 자료들을 분석하고자 하는데 목적이 있다. Nelder 와 Wedderburn(1972)이 일반화 선형모형을 개발하고 McCullagh 과 Nelder(1989, 2nd edition)가 1983년 이에 대한 소개 책자를 발간한 후 통계학 분야에서 널리 사용되고 있다.

3.1. 반응변수의 분포와 분산함수

일반화 선형모형은 로짓, 프라빗, 로그선형모형등을 포함하는데, 반응변수의 분포를 정규분포뿐 아니라 포아송분포, 이항분포, 감마분포, 음이항(negative binomial)분포, 역정규(inverse Gaussian)분포 등을 포함하는 일모수지수족(one-parameter exponential family)으로 확장하였다. 즉 정규분포를 따르는 연속형 자료의 회귀분석 뿐만 아니라, 결점수등의 개수 자료는 포아송분포, 불량율 등의 비율에 관한 자료는 이항분포, 양의 값을 갖는 연속변수로 등변동계수(constant coefficient of variation)를 갖는 자료는 감마분포를 이용하여 회귀분석할 수 있다.

일반화 선형모형에서는 평균과 분산이 다음과 같이 표현된다.

$$E(y) = \mu, \quad Var(y) = \phi V(\mu) \quad (1)$$

여기서, ϕ 를 산포모수(dispersion parameter)라고 하며 $V(\mu)$를 분산함수라고 부른다. 산포모수 ϕ는 반응변수 y의 분산 중 y의 평균에 상관없는 부분을 나타내며 분산함수 $V(\mu)$ 는 y의 분산이 어떻게 y의 평균에 따라 변하는 가를 나타낸다.

분포에 따라 분산함수들이 달라지는 바 이를 <표 1> 로 요약하였다. 분산함수가 정해지면 일반화 선형모형의 지수족 중에서 분포가 지정되므로 일반화 선형모형에서는 분포의 지정과 분산함수의 지정은 동치이다. 정규분포의 경우는 $V(\mu)$=1 로 분산이 평균에 따라 변하지 않음, 즉 등분산성을 나타낸다. 그러므로 분산함수의 지정은 등분산성 가정의 확장이라고 볼 수 있다.

예를 들면 포아송분포의 경우 $V(\mu)=\mu$ 로 분산이 평균이 증가함에 따라 증가함을 나타낸다. 이항분포에서는 p가 모비율을 나타내며 μ =np의 관계가 있다. 그러므로 이항분포에서는 $V(\mu) = \mu(n-\mu)/n$ = np(1-p)로 p가 0 이나 1 에 가까우면 분산이 작아짐을 알 수 있다. 감마분포에서는 $Var(y)=\phi\mu^2$로 $\phi^{1/2}$는 변동계수를 나타낸다. 분산함수가 주어지면 일반화 선형모형에서는 회귀계수들이 가중최소제곱법(weighted least square)을 이용한 최우 추정법(maximum likelihood estimation)으로 추정되므로 분산함수의 올바른 지정이 회귀계수 추정치들의 효율성에 큰 영향을 미친다. 예를 들어 이항분포를 따르는 비율 자료의 분석을 등분산성을 가정한 회귀분석으로 한다면 매우 비효율적인 분석이 된다.

<표 1> 분포에 따른 분산함수

분포	산포모수	분산함수
정규분포	σ^2	1
포아송분포	1	μ
이항분포	1	$\mu(n-\mu)/n$
음이항분포	1	$\mu + \alpha\mu^2$
감마분포	ϕ	μ^2
역정규분포	ϕ	μ^3

3.2. 연관함수와 가법성

회귀분석 모형에서는 가법성이 다음과 같이

$$\mu = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k$$

평균(μ)척도에서 성립한다고 가정하는 반면, 일반화 선형모형에서는 가법성이 다음과 같이 연관함수 $g(\ )$를 통해

$$\eta = g(\mu) \quad \cdots\cdots\cdots\cdots \quad (2)$$
$$= \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k$$

성립한다고 한다. 분포에 따라 자주 사용되는 연관함수들을 <표 2> 에 정리하였다.

<표 2> 분포와 연관함수

분포	연관함수	모형
정규분포	μ	Linear Model
이항분포	$\log p/(1-p)$ $\Phi^{-1}(p)$	Logit Model
포아송분포	$\log \mu$	Probit Model
음이항분포	$\log \mu$	Log-linear Model
감마분포	$\log \mu$ 또는 $1/\mu$	
역정규분포	$\log \mu$ 또는 $1/\mu$	

일반화 선형모형에서는 자료의 변환을 통해 가법성을 성취하는 것이 아니라 모수의 변환을 통해 가법성을 충족시킨다. 예를 들어 개수 자료들을 분석하는데 사용되는 포아송분포에서 가법성을 성취하기 위해 로그변환이 필요하다 하자. 이 경우 $\log y$ 변환자료에 대한 회귀분석은 연관함수를 $\log\mu$로 잡고 일반화 선형모형을 사용하는 것과 유사하다. 그러나 로그값이 0 값에서 정의되지 않으므로 자료변환 $\log y$ 에서는 자료의 손실이 발생할 수 있으나 모수변환 $\log\mu$ 를 통해 가법성을 가정하는 것은 아무런 자료의 손실을 초래하지 않는다. 예를 들어 평균 μ = 1 인 포아송분포를 생각해보자. 이 경우 반응변수 y가 0 일 확률이 0.3679로 $\log\mu$ = 0 은 항상 정의되나 $\log y$ 는 약 37%의 자료에서 정의되지 않는다.

연관함수의 중요한 역할로 다음의 두 가지를 들 수 있다. 포아송 경우를 생각해 보면 평균 μ 는 항상 양수이나 $\eta(=X\beta)$ 가 취할 수 있는 값의 범위는 $(-\infty,\infty)$ 이다. 그러므로 정규분포를 가정한 회귀분석에서는 $\eta=g(\mu)=\mu$ 로 잡

기 때문에 평균 μ의 예측치가 음의 값을 가질 수 있으나 로그 연관함수를 가정하면 $\mu=\exp(\eta)=\exp(X\beta)$의 예측치가 항상 양의 값만을 취한다.

이제 두 수준을 갖는 A, B 두 인자를 가정하고 다음의 가법모형을 생각해 보자.

$$\mu_{ij} = \beta_0 + a_i + b_j . \qquad (3)$$

<표 3>의 자료 1에서는 $\beta_0=25$, $a_1=-10$, $a_2=10$, $b_1=-5$, $b_2=5$ 로 (3)의 관계를 만족하나 자료 2는 (3)과 같이 A, B 두 인자의 주효과만의 가법모형으로는 나타낼 수 없고 A, B의 교호작용이 필요하다. 그러나 자료 2는 곱 모형으로 설명되므로 로그 연관함수를 사용하면 다음과 같이 표현된다.

$$\log \mu_{ij} = \beta_0 + a_i + b_j .$$

이 경우 자료 2는 $\beta_0=0.5\log 600$, $a_1=-0.5\log 3$, $a_2=0.5\log 3$, $b_1=-0.5\log 2$, $b_2=0.5\log 2$ 의 관계를 만족한다. 그러므로 연관함수의 적절한 선택은 요인실험법(factorial design)에서 불필요한 교호작용을 없애줄 수 있다.

<표 3> A, B 인자수준에 따른 μ 값의 변화

자료1

		B	
		1	2
A	1	10	20
	2	30	40

자료 2

		B	
		1	2
A	1	10	20
	2	30	60

4. 다구찌 소음신호

다구찌(1985, 1987) 품질개선실험의 목표는 제품간의 변이를 줄이고 제품의 평균을 성능목표치에 맞추는 적정 공정과정을 찾아내는데 있다. 다구찌(1987)는

$$SN = 10 \log \frac{\mu^2}{Var(y)}$$

을 정의하고 이의 추정치로써 다음과 같은 SN 비를 사용할 것을 주장하였다.

$$SN\text{ 비} = 10 \log \frac{\bar{y}^2}{s_y^2} .$$

다구찌의 SN 비는 제품간 변이를 반영하는 측도로서 SN 비가 크면, 변이가 작다는 것을 의미한다. 그러므로, 다구찌는 SN 비를 크게 하는 공정과정 조합을 품질개선 실험을 통해 찾아야한다고 하였다.

감마분포에서는 $Var(y)=\phi\mu^2$ 이므로

$$SN = -10 \log \phi$$

가 된다. 그러므로 SN 비를 크게 하는 것은 산포모수 ϕ를 작게 하는 것과 동치이다. 일반적으로 일반화 선형모형에서는 $E(y) = \mu$, $Var(y) = \phi V(\mu)$ 로서 ϕ는 반응변수 y의 분산 중 평균 μ와 상관없는 부분을 나타낸다. 그러므로 SN 비를 크게 하는 것은 ϕ를 줄임으로써 제품의 평균 μ에 영향을 주지않고 제품간 변이를 줄이고자 하는 것이다. Nair 와 Pregibon(1986), Leon 과 그외(1987) 등은 ϕ의 함수의 추정치를 PerMIA(Performance Measure Independent of Adjustment)라고 하였다. 그러므로 다구찌의 SN 비는 분산함수가 $V(\mu) = \mu^2$인 경우에 한하여 PerMIA가 된다.

일반화 선형모형에서 $Var(y) = \phi V(\mu)$ 로 정의하는 바 이 식은 일반적인 분산함수에서 PerMIA가 정의될 수 있게 해준다. 또한 회귀분석법을 통해 SN 비를 크게 하는 공정과정을 찾는다는 것은 가법성이 $\log\phi$의 측도에서 만족됨을 의미한다. 즉, 다구찌의 SN 비를 통한 품질개선 실험은 일반화 선형 모형에서 보면,

(i) $V(\mu) = \mu^2$와

(ii) 가법성이 $\log\phi$의 측도에서 만족된다는

두 가지 가정을 한다고 볼 수 있다. 한편, Box (1988) 는 z = log y 의 자료변환이 E(z) ≈ log μ 와 Var(z) ≈ ϕ 의 관계를 만족하므로 변환

된 자료 z 에 대한 회귀분석을 통해 다구찌 품질 개선실험자료를 분석하여야 한다고 주장하였다.

5. 일반화 선형모형의 확장

일반화 선형모형은 산포모수 ϕ 가 상수인 것을 가정한다. 제품간 변이를 줄이고 제품평균을 성능목표치에 맞추어 적정 공정과정을 찾는 다구찌 품질개선 실험자료 분석을 위하여 Nelder와 Lee(1991)는 다음과 같은 평균과 분산의 동시 모형을 통해 일반화 선형모형을 확장하였다.

$$\eta = g(\mu) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k$$

$$\log \phi = \gamma_0 + Z_1\gamma_1 + Z_2\gamma_2 + \cdots + Z_p\gamma_p \quad \cdots\cdots\cdots\cdots (4)$$

$(X_1, X_2, \cdots, X_k)$와 $(Z_1, Z_2, \cdots, Z_p)$의 두 군데 모두 들어가는 설명변수는 평균과 분산에 동시적으로 영향을 주며 $(X_1, X_2, \cdots, X_k)$에만 들어가는 변수는 평균에만, $(Z_1, Z_2, \cdots, Z_p)$에 들어가는 변수는 분산에만 영향을 주는 공정과정들이 된다. 그러므로 평균과 분산에 동시적으로, 또는 분산에만 영향을 주는 공정과정들을 통해 제품간 변이를 줄인 후 평균만 영향을 주는 공정과정을 통해 제품의 평균을 성능 목표치에 맞춘다. 평균과 분산의 동시 모형(4)에서 사용할 연관함수 $g()$로 표2를 이용할 수 있겠다. 즉 반응변수가 이항분포를 따르면 로짓 또는 프라빗등을 사용하고 포아송분포를 따르면 로그함수등을 이용할 수 있겠다. 이 모형은 Lee와 Nelder (1998)와 Nelder와 Lee (1998)에서 확장되었고 이영조 (1993)에의해 국내에 소개되었다.

본 발표에서는 일반화 선형모형의 다구찌 품질개선 실험자료 분석에서의 확장뿐만 아니라 그 외에 필요한 확장들에 대하여 자료분석을 통해 심도있게 논의하고자 한다.

참 고 문 헌

이영조(1993). 다구찌 실험 분석에 있어서 일반화선형모형 대 자료변환, 응용통계 연구, 6권, 2호, 253-263.

Box, G. E. P. (1988), "Signal-to-Noise Ratios, Performance Criteria and Transformations," Technometrics, 30, 1-17.

Box, G. E. P. and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), Journal of the Royal Statistical Society, B, 26, 211-252.

Hougaard, P. (1982), "Parametrizations of Non-Linear Models," Journal of the Royal Statistical Society, B, 44, 244-252.

Lee, Y. and Nelder, J. A. (1998), "Generalized Linear Models for the Analysis of Quality-Improvement Experiments." The Canadian Journal olf Statistics, 26, 95-105.

Leon, R. V., Shoemaker, A. C., and Kackar, R. N. (1987), "Performance Measures Independent of Adjustment" (with discussion), Technometrics, 29, 253-285.

McCullagh, P. and Nelder, J. A. (1989), Generalized Linear Models, 2nd ed. London: Chapman and Hall.

Nair, V. N. and Pregibon, D. (1986), "A Data Analysis Strategy for Quality Engineering Experiments," AT&T Technical Journal, 65, 73-84.

Nelder, J. A. and Lee, Y. (1991), "Generalized Linear Models for the Analysis of Taguchi-Type Experiments." Applied Stochastic Models and Data Analysis, 7, 107- 120.

Nelder, J. A. and Lee, Y. (1992) "Likelihood, Quasi-likelihood and Pseudo-likelihood: Some Comparisons." Journal of the Royal Statistical Society, B, 54,

Nelder, J. A. and Lee, Y. (1998) "Letter to Editor." Technometrics, 40, 168-175.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," Journal of the Royal Statistical Society, A, 132, 370-384.

Taguchi, G. (1985), "Quality Engineering in Japan," Communications in Statistics, A, 14, 2785-2801.

Taguchi, G. (1987), System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost, White Plains, NY: UNIPUB/ Krau International.