응용통계연구 제 17권 3호, 2004년, pp. 489–498

분포함수를 기초로 일반화가중선형모형*

여인권[1)]

요 약

이 논문에서는 일반화가중선형모형이라는 새로운 형태의 선형모형을 제시한다. 일반화가중선형모형은 설명변수와 반응변수의 관계를 설명분포함수의 선형결합이 반응변수의 평균에 대한 연결분포함수를 통해 모형화 되는 형태를 가지는 것으로 가정한다. 이 모형은 일반화선형모형에서 연결함수를 선택할 때 발생할 수 있는 모수공간과 선형예측값의 공간이 일치하지 않을 수 있다는 문제가 발생하지 않고 모수에 대한 해석이 용이하다는 장점이 있다. 이 논문에서는 설명분포함수와 연결분포함수를 선택하는데 있어 발생할 수 있는 문제와 해결책에 대해 알아본다. 또한 모형에 포함되어 있는 모수를 추정하는데 고려해야 할 주의 사항과 이 사항들을 고려한 최대가능도추정법과 재표집 방법을 이용한 구간추정과 가설검정에 대해 알아본다.

주요용어: 공액분포, 모수적 변환, 베타 혼합체, 붓스트랩, 지수족

1. 서론

고전적 선형모형에서는 모수에 대한 통계적 추론을 하기 위해 오차들은 서로 독립이며 평균이 0이고 분산이 같은 정규분포를 따른다고 가정한다. 그러나 자료가 이 가정들을 만족하지 않거나 자료의 특성상 정규분포를 가정할 수 없는 경우를 종종 보게 된다. 이런 문제를 해결하기 위한 방법으로 변수변환이 이용되고 있는데 이진반응자료처럼 정규분포와 유사한 형태로 바꾸어 주는 변환을 찾기 어려운 경우도 있고 적절한 형태의 변환이 존재하더라도 사용 목적에 따라 다른 형태의 변환을 선택해야 하는 경우가 있다. 예를 들어, 확률변수 Y가 포아송분포를 따른다고 할 때 분산안정화를 시키기 위해서는 $\sqrt{Y}$가 사용되는데 반하여 대칭을 이루게 하기 위해서는 $Y^{2/3}$이 적절하다는 것이 알려져 있다 (McCullagh와 Nelder(1989), p 22). 이런 경우 어디에 초점을 맞추는가에 따라 다른 형태의 변환을 선택해야 한다는 단점이 있다.

Nelder와 Wedderburn(1972)는 정규분포뿐만 아니라 이항분포, 포아송분포, 감마분포를 포함한 지수족(exponential family)을 근거로 만들어진 일반화선형모형(generalized linear models)을 소개하여 선형모형에서의 등분산성과 정규성의 문제를 해결하였다. 반응확률변수 $Y_1, \ldots, Y_n$은 서로 독립이며 평균이 $E(Y_i) = \mu_i$인 지수족을 따르고 반응평균 μ_i는 설명변수 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$에 의해 영향을 받는다고 하자. 선형모형에서는 설명변수와 반응

* 이 논문은 2002년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2002-003-C00028)

1) (140-742) 서울특별시 용산구 청파동 2가 숙명여자대학교 이과대학 수학통계학부 조교수
E-mail: inkwon@sookmyung.ac.kr

평균 간의 관계를 미지의 모수 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$에 대해 $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$로 표시하는 반면 일반화선형모형에서는 미분 가능하고 단조함수인 $g(\cdot)$에 대해

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

를 만족한다고 가정한다. 이 때 반응평균 μ_i와 선형예측값(linear predictor) $\boldsymbol{x}_i^T\boldsymbol{\beta}$를 연결시키는 함수 $g(\cdot)$를 연결함수(link function)라고 한다. 이진반응자료를 설명할 때 많이 사용되는 로짓함수(logit function), 프로빗함수(probit function), 보로그로그(complementary log-log) 등이 대표적인 연결함수이다. 연결함수가 항등함수이고 Y의 분포가 정규분포이면 일반화 선형모형은 선형모형과 일치한다.

일반적으로 연결함수를 어떻게 설정하는가에 따라 분석 결과에 차이가 있을 수 있기 때문에 연결함수를 결정하는 과정은 자료 분석에서 매우 중요한 단계이다. 대부분의 분포에서 정준연결함수(canonical link function)가 추론에 있어 간단하면서도 이론적으로 좋은 성질들이 있어 호감이 가는 함수이지만 감마분포에서의 $g(\mu) = 1/\mu$와 같이 항상 적절한 구조를 제공하는 것은 아니다. 대부분의 경우 연결함수는 평균의 모수공간을 고려하여 선택하는데 예를 들어 포아송분포의 경우 평균은 양수이기 때문에 어떤 설명변수에 대해 $\mu \leq 0$을 가지는 연결함수는 적절한 함수라고 할 수 없다. 비록 포아송모형에 대해 로그함수는 정준연결함수이고 실제 자료 분석에서 가장 많이 사용되고 있지만 자료적합의 관점에서 볼 때 로그함수보다 선형예측값과 반응평균간의 관계를 잘 설명할 수 있는 다른 형태의 연결함수가 있을 수 있다. 또한 모든 설명변수의 효과가 반응평균의 로그값으로만 표시되기 때문에 어떤 설명변수가 반응평균에 다른 척도로 영향을 주는 경우 이 설명변수에 대응하는 회귀계수의 유의성이 떨어질 수도 있다.

이 논문에서는 일반화선형모형과 같이 지수족을 근거로 하고 있으면서 설명변수와 반응평균의 관계를 다양한 구조로 설명할 수 있는 모형을 제시하고 이 모형에 포함되어 있는 모수에 대한 추론 방법에 대해 알아본다. 특히 모수에 따라 다양한 형태를 제공하는 베타분포의 혼합체를 이용함으로써 설명변수와 반응평균의 관계가 단조이면서 비선형적일 경우에도 모형화할 수 있는 특징을 가지고 있다.

2. 일반화가중선형모형

확률변수 $Y_1, \ldots, Y_n$는 서로 독립이고 밀도함수가 다음과 같은 지수족을 따른다고 하자.

$$f(y_i;\theta_i) = \exp\left[a^{-1}(\phi_i)\left\{y_i\theta_i - b(\theta_i)\right\} + c(y_i,\phi_i)\right].$$

여기서 함수 $a(\cdot)$, $b(\cdot)$, 그리고 $c(\cdot)$의 형태는 알려져 있다고 가정한다. 모수 ϕ_i가 알려져 있을 때 밀도함수는 정준모수(canonical parameter)가 θ_i인 지수족이 되고 일반적인 경우 Y_i의 평균과 분산은 각각 $E(Y_i) = \mu_i = b'(\theta_i)$와 $var(Y_i) = a(\phi_i)b''(\theta_i)$로 구할 수 있다. 이 논문에서는 편의상 모수 ϕ_i는 알려진 상수로 취급한다. 일반화선형모형의 설정에서는 다음의 관계식이 성립한다.

$$g(b'(\theta_i)) = g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

이 논문에서 연구하고자 하는 모형은 평균 μ_i와 설명변수 $\boldsymbol{x}_i$의 관계를 다음과 같이 가정한다.

$$F_\mu(\mu_i) = \sum_{j=1}^{p} \alpha_j F_j(x_{ij})^{\beta_j} \{1 - F_j(x_{ij})\}^{1-\beta_j}. \tag{2.1}$$

여기서, $j = 1, \ldots, p$에 대해, $\alpha_j \geq 0$이고 $\sum_{j=1}^{p} \alpha_j = 1$를 만족하며 β_j는 0 아니면 1의 값을 가진다. 함수 F_μ와 F_j들은 임의의 분포함수로 F_μ를 연결분포함수(link distribution function)라고 부르고 함수 F_j를 설명분포함수(explanatory distribution function)라고 부른다. 관계식 (2.1)에서 보는 것처럼 반응평균의 연결분포가 설명변수들의 설명분포의 가중합으로 표시되기 때문에 이 모형을 일반화가중선형모형이라고 부른다. 이 모형의 장점은 앞절에서 언급한 것과 같이 일반화선형모형에서 연결함수의 값과 선형예측값의 영역이 일치하지 않을 때 발생할 수 있는 문제가 없다는 것이다.

각각의 설명변수 x_{ij}는 F_j를 통해 0과 1사이의 값으로 표준화되기 때문에 모수 α_j은 모형에 있어 j번째 설명변수의 상대적 중요도를 나타내는 값으로 해석될 수 있다. 또한 상대적으로 큰 값을 가지는 변수를 모형에 추가하고 작은 값을 가지는 변수를 제거하는 모형선택의 기준으로 사용될 수 있다. 모수 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$를 가중모수(weight parameter)라고 부른다. 연결함수 F_μ와 설명함수 F_j는 증가함수이므로 모수 β_j가 1이면 j번째 설명변수가 평균에 대해 양의 효과를 가지는 것을 의미하고 0이면 음의 효과를 가지는 것을 의미하기 때문에 모수 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$를 효과모수(effect parameters)라고 부른다.

일반화가중선형모형은 연결분포와 설명분포의 선택에 따라 기존의 일반화선형모형을 포함하거나 근사시킬 수 있다.

예제 2.1: 확률변수 Y는 성공확률이 μ인 베르누이분포를 따른다고 하자. 연결분포 $F_\mu(\cdot)$를 0과 1사이의 균일분포로 정하고 설명분포 $F_1(\cdot)$를 위치모수가 γ_0이고 척도모수가 γ_1인 로지스틱분포로 정하면 관계식 (2.1)는 다음과 같이 쓸 수 있다.

$$\gamma_0 + \gamma_1 x = g(\mu) = \begin{cases} \log\{\mu/(1-\mu)\}, & \beta_1 = 1 \\ \log\{(1-\mu)/\mu\}, & \beta_1 = 0. \end{cases}$$

이 모형은 이진반응자료에 대한 로지스틱 회귀모형을 포함한다.

예제 2.2: 확률변수 Y는 평균이 $\mu > 0$인 지수족을 따른다고 하자. 설명분포함수 $F_1(\cdot) = F_{\gamma_0,\gamma_1}(\cdot)$는 위치모수가 γ_0이고 척도모수가 γ_1인 임의의 분포함수라고 할 때 연결분포함수를 $F_\mu(\cdot) = F_{0,1}(\log(\cdot))$로 설정하면

$$\gamma_0 + \gamma_1 x = g(\mu) = \begin{cases} \log(\mu), & \beta_1 = 1 \\ F_{0,1}^{-1}\left[1 - F_{0,1}\{\log(\mu)\}\right], & \beta_1 = 0. \end{cases}$$

3. 분포선택

3.1. 연결분포함수선택

이 모형을 실제자료에 적용시키기 위해서는 분포함수 F_μ와 F_j들을 사전에 지정해야

한다. 적절한 연결분포함수 F_μ를 지정하는데 있어 고려할 점은 F_μ에 대응하는 밀도함수의 지지(support)와 μ에 대한 모수공간 Ω가 일치하는가와 자료를 얼마나 잘 적합 시키는가이다. 일단 모수공간과 지지가 일치하는 모든 분포가 F_μ로 사용될 수 있으나 반응변수의 분포가 지수족에 속한다는 것을 고려하면 Pitman-Koopman 보조정리(참고, Lehman과 Casella(1998))에 의해 공액분포(conjugate distribution)가 존재하고 이 분포의 지지가 모수공간과 일치하기 때문에 특별한 대안이 없는 경우 공액분포를 연결함수로 사용할 수 있다. 또한 다양한 형태를 가지는 관계식을 설명하기 위해 일반화선형모형에서 Aranda-Ordaz (1981), Stukel (1988), Taylor (1988), Cheng과 Wu (1994) 등이 했던 것처럼 모수적 변환을 이용하여 자료에 대한 적합력을 높일 수도 있다.

이 논문에서는 모수를 하나만 포함하고 있는 두 가지 형태의 모수적 변환을 고려하였다. 모수 λ를 변환의 형태를 결정하는 변환모수라고 할 때 첫 번째 변환 $\psi(v,\lambda)$은 정의역과 치역이 모두 [0,1]인 임의의 증가함수이다. 이와 같은 성질을 만족하는 함수로, $\lambda > 0$에 대해, $\psi(v,\lambda) = v^\lambda$ 또는, 임의의 상수 k에 대해, 베타분포의 분포함수 $\mathrm{IB}(v;\lambda,k)$ 또는 $\mathrm{IB}(v;k,\lambda)$ 등을 생각할 수 있다. 이런 함수를 고려한 경우 연결분포함수는 다음과 같이 쓸 수 있다.

$$F_\mu(u) = \psi^{-1}\{F(u),\lambda\}. \tag{3.1}$$

여기서 F는 앞에서 언급한 공액분포일 수도 있고 자료적합에 직접적인 관련이 없는 임의의 분포함수도 될 수 있다.

두 번째 형태의 연결분포는 $F_\mu(u) = F\{\phi(u,\lambda)\}$의 구조를 가지는데 여기서 변환 $\phi(u,\lambda)$는 치역과 정의역이 모두 Ω인 증가함수이다. 만약 $\Omega = R^+$이거나 $\Omega = R$라면 John과 Draper (1980)의 율변환(modulus transformation)이나 Yeo과 Johnson (2000)의 확장된 멱변환 등이 사용될 수 있다. 율변환의 경우 u의 부호에 따라 한쪽이 볼록(convex)이면 다른쪽은 오목(concave)이 되어 $\Omega = R$에서 한쪽에서 과소추정(under-estimation)되고 다른 한쪽에서 과대추정(over-estimation)되는 경우에 효과적으로 사용될 수 있다. 이에 반하여 Yeo-Johnson 변환은 $\lambda < 1$일 때 전 구간에서 오목이고 $\lambda > 1$일 때 볼록이 되어 양쪽 꼬리에서 과소추정되거나 아니면 과대추정되는 경우 효과적이다. $\Omega = [0,1]$인 경우에는 위에서 언급한 ψ가 사용될 수 있다.

3.2. 설명분포함수선택

함수 $F_j^*(\cdot)$를 (2.1)의 관계를 설명하기 위한 최적의 설명분포함수라고 하자. 대부분의 자료 분석에서는 $F_j^*(\cdot)$를 알 수가 없기 때문에 자료들 간의 관계를 고려하여 추정하거나 선택해야 한다. 미지의 증가함수를 모형화 하거나 추정하기 위해 베타분포함수의 혼합체(mixture)를 이용하는 방법이 Mallick와 Gelfand(1994)등의 논문들에 소개되고 있는데 이 논문에서도 함수 $F_j^*(\cdot)$를 추정 또는 선택하기 위해 베타혼합체를 이용한다.

함수 $F_j^0(\cdot)$를 $F_j^*(\cdot)$를 추정하기 위해 선택된 임의의 중심분포함수라고 하자. 이 중심분포를 선택할 때 해당 설명변수의 분포적 특성을 고려하면 확률적분변환에 의해 F_j^0는 균일분포를 따르는 형태가 되어 설명변수의 효과를 설명하는데 있어 적절한 해석이 가능해질 수 있다. 설명변수의 분포적 특징을 모르는 경우 F_j^0를 일정구간에서의 균일분포라고 지정

한다면 분석자가 분포를 임의로 선택함으로써 발생할 수 있는 객관성의 결여문제를 제거할 수도 있다. 베타분포함수 $\mathrm{IB}(F_j^0(\cdot);c,d)$는 형태모수 c와 d에 따라 다양한 형태를 제공하기 때문에 $F_j^*(\cdot)$는 다음과 같이 근사될 수 있다.

$$F_j^*(u) \simeq \sum_{l=1}^{L_j} \omega_l \mathrm{IB}(F_j^0(u);c_l,d_l).$$

여기서 L_j는 베타혼합체에 포함된 베타분포함수의 개수이고 ω_l들은 혼합 가중값으로 $\omega_l \geq 0$와 $\sum_{l=1}^{L_j}\omega_l = 1$를 만족한다.

위의 베타혼합체를 사용하는데 있어 발생하는 문제는 추정하거나 결정해야 할 모수가 $\omega = (\omega_1,\ldots,\omega_L)^T$, $\boldsymbol{c} = (c_1,\ldots,c_L)^T$, 그리고 $\boldsymbol{d} = (d_1,\ldots,d_L)^T$로 L이 크면 모수가 많아져 분포선택에 부담이 될 수 있다는 것이다. 이 문제를 효과적으로 해결하기 위한 방법으로는 베타분포의 형태모수를 $c_l = \sigma l$와 $d_l = \sigma(L+1-l)$로 정하는 것이다. 이렇게 하면 다양한 형태의 베타분포를 표시할 수 있으면서 추정 또는 지정해야 할 모수가 하나로 줄어드는 효과가 있다. 또한 $\boldsymbol{\omega}$를 어떤 한 l에서만 1의 값을 가지고 나머지는 0이 되는 것으로 가정하면 L개의 베타분포함수 중에서 가장 적절하다고 판단되는 분포함수 하나를 선택하는 문제로 바꿀 수 있다.

4. 모형추정과 가설검정

모수 $\boldsymbol{\gamma}$는 연결분포함수와 설명분포함수에 포함되어 있는 미지의 모수라고 하자. 표기상의 편이를 위해 $\boldsymbol{F}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{x}_i) = \sum_{j=1}^{p}\alpha_j F_j(x_{ij};\boldsymbol{\gamma}_j)^{\beta_j}\left\{1-F_j(x_{ij};\boldsymbol{\gamma}_j)\right\}^{1-\beta_j}$라고 쓴다. 일반적으로 반응평균 μ의 공간은 연속형이기 때문에 연결분포함수 F_μ는 연속형으로 가정한다. 관계식 (2.1)으로부터 $\mu_i = F_\mu^{-1}\{\boldsymbol{F}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{x}_i)\}$로 쓸 수 있고 방정식 $h^{-1}(\theta_i) = \mu_i = b'(\theta_i)$를 이용하면

$$\theta_i = h(\mu_i) = h\left[F_\mu^{-1}\{\boldsymbol{F}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{x}_i)\}\right] = h(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{x}_i)$$

이고 로그가능도함수는 다음과 같이 쓸 수 있다.

$$l_n(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{y},\mathbf{X}) = \sum_{i=1}^{n} a^{-1}(\phi_i)\left\{y_i h(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{x}_i) - b(h(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{x}_i))\right\}. \tag{4.1}$$

여기서 $\mathbf{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$는 계획행렬(design matrix)을 나타낸다.

일반화가중선형모형에서는, $i=1,\ldots,p$에 대해, $\alpha_i \geq 0$, $\sum_{i=1}^{p}\alpha_i = 1$, 그리고 β_i는 0 아니면 1이어야 한다는 모수공간 상의 제약조건 때문에 최대가능도추정량을 구하기 위해 일반적으로 사용되는 점수벡터(score vector)를 사용할 수 없다. 최대가능도추정량을 얻는 방법은 모든 가능한 $\boldsymbol{\beta}$의 조합에 대해 라그랑지 배수법을 사용하여 모수를 추정한 후 로그가능도함수 (4.1)를 최대로 만드는 $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$의 추정량을 선택한다. 문제는 설명변수가 많아지면 추정해야 할 모수가 많아져 수치해석학적으로도 최대가능도추정값을 계산하기 어렵다는 것이다. 이 경우에는 simulated annealing과 같은 확률적 탐색방법을 사용하는 것이 효과

적일 수 있다. 또 다른 문제는 모수공간의 제약조건 때문에 최대가능도추정량의 점근적 성질을 유도하기 어렵다. 이 논문에서는 붓스트랩(bootstrap)과 같은 재표집(resampling) 방법을 이용하여 모수에 대한 신뢰구간 및 가설검정에 대해 알아본다.

선형모형에서의 붓스트랩 방법처럼 일반화가중선형모형에서도 bootstrapping pairs 방법과 bootstrapping residuals 방법이 사용될 수 있다. Boostrapping pairs 방법은 쌍으로 이루어진 원래 자료의 집합 $\{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\}$에서 n개의 쌍을 랜덤하게 복원추출하여 얻은 붓스트랩 표본 $(y_1^*, \boldsymbol{x}_1^*), (y_2^*, \boldsymbol{x}_2^*), \ldots, (y_n^*, \boldsymbol{x}_n^*)$를 이용한다. 이 방법은 사용이 표본을 쉽게 추출할 수 있다는 장점이 있지만 자료에서 빠지거나 반복된 자료가 나타나기 때문에 재표본의 특성이 원래 자료의 특성과 차이가 있을 수 있다. Boostrapping residual 방법을 사용하기 위해서는 일단 모수 $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$의 추정값 $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$를 구하고 $\hat{\mu}_i = F_\mu^{-1}\{\boldsymbol{F}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \boldsymbol{x}_i)\}$를 계산해야 한다. 동일한 분산을 가지는 정규분포를 근거한 일반선형모형과 다르게 일반화가중선형모형은 지수족을 근거로 하기 때문에 boostrapping residual 방법에서는 잔차 $r_i = y_i - \hat{\mu}_i$를 사용하기보다는 가정한 분포에 따라 Pearson 잔차, Anscombe 잔차 또는 deviance 잔차 등을 사용하는 것 더 효율적일 수 있으며 붓스트랩 과정이 일치성을 만족하는 이론적 토대가 될 수 있다. 잔차 $\tilde{r}_i$를 분포의 형태에 따라 선택된 잔차라고 하면 잔차의 합이 0이 되도록 만든 표준화 잔차 $\tilde{r}_i^s = \tilde{r}_i - n^{-1}\sum_{j=1}^n \tilde{r}_j$를 구한다. 붓스트랩 잔차 $r_1^*, \ldots, r_n^*$은 $\{\tilde{r}_1^s, \ldots, \tilde{r}_n^s\}$에서 랜덤하게 복원추출되고 붓스트랩 표본은 다음과 같이 얻을 수 있다.

$$y_i^* = \hat{\mu}_i + \sqrt{v_i} r_i^*, \quad i = 1, \ldots, n.$$

여기서 v_i는 분산함수 $b''(\theta_i)$의 추정량을 의미한다.

일반적으로 어떤 설명변수의 유의성은 그 설명변수를 포함한 모형의 적합도와 포함하지 않은 모형의 적합도 간의 차이가 어떤 기준보다 큰지 아닌지를 보고 판단한다. 예를 들어 p번째 설명변수가 모형에서 제외된 설명분포의 가중합을

$$\boldsymbol{F}(\boldsymbol{\alpha}^R, \boldsymbol{\beta}^R, \boldsymbol{\gamma}^R, \boldsymbol{x}_i) = \sum_{j=1}^{p-1} \alpha_j^R F_j(x_{ij}; \boldsymbol{\gamma}_j^R)^{\beta_j^R} \left\{1 - F_j(x_{ij}; \boldsymbol{\gamma}_j^R)\right\}^{1-\beta_j^R}$$

라고 하자. 여기서 $\sum_{j=1}^{p-1} \alpha_j^R = 1$이고 $\boldsymbol{\gamma}^R$은 $\boldsymbol{\gamma}$에서 p번째 설명분포함수에 포함된 모수를 제외시킨 모수벡터를 의미한다. 그러면 p번째 변수의 유의성에 대한 가설은 $H_0 : \alpha_p = 0$와 $H_1 : \alpha_p > 0$로 표시할 수 있으며 모형으로는 다음과 같은 가설을 검정하는 것으로 생각할 수 있다.

$$H_0 : \ \mu_i = \boldsymbol{F}(\boldsymbol{\alpha}^R, \boldsymbol{\beta}^R, \boldsymbol{\gamma}^R, \boldsymbol{x}_i), \qquad H_1 : \ \mu_i = \boldsymbol{F}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}_i).$$

검정통계량 T는 Pearson의 X^2 통계량과 같이 모형의 적합도를 나타내는 통계량이라고 하자. 대립가설 하에서의 적합도 $T(\hat{\boldsymbol{\mu}}_1)$가 귀무가설 하에서의 적합도 $T(\hat{\boldsymbol{\mu}}_0)$ 보다 상대적으로 작으면 귀무가설을 기각시킬 수 있을 것이다. 문제는 기각역이나 유의확률(p-value)를 구하기 위해서는 귀무가설 하에서 적합도 $T(\hat{\boldsymbol{\mu}}_0)$의 분포를 유도해야하는데 위 모형에서는 쉽지 않다는 것이다. 이 경우 앞에서 언급한 bootstrapping pairs나 bootstrapping residuals과 같은 재표집 방법을 이용하여 근사시킬 수 있다. Boostrapping pairs에서는 원 자료에서 랜덤하게 복원추출한 재표본을 이용하여 귀무가설 하에서의 모형을 적합시킨 후 적합도를 계

산하는 것으로 $T(\hat{\mu}_0)$의 분포를 근사할 수 있다. Bootstrapping residuals에서는 원 자료를 귀무모형에 적합시킨 후 나온 잔차를 이용하여 재표본을 추출하여 다시 귀무모형에 적합시켜 적합도를 계산하는 방법으로 분포를 근사시킨다. 연결분포함수와 설명분포함수에 포함된 모수에 대해서도 유사한 방법을 적용하여 검정할 수 있다.

5. 예제

Milicer와 Szczoka(1966)는 1965년 Warsaw지역에 거주한 9.21세부터 17.58까지의 3918명 소녀들을 대상으로 초경연령에 대한 자료를 분석하였다. Aranda-Oradaz (1981)는 이 자료를 분석하는데 프로빗 연결함수가 적절하다고 주장하였으나 Yeo(2001)에 의하면 로지스틱, 프로빗, 보로그로그 연결함수 모두 자료적합에 있어 개선의 여지가 있는 것으로 나타났다. Guerrero와 Johnson (1982)는 설명변수를 Box-Cox 변환시켜 정규분포에 가깝게 변형시킨 후 프로빗 모형을 적용시키는 방법을 사용하여 기존의 분석결과를 대폭 개선시켰다.

이 자료를 제안된 모형으로 분석하기 위한 사전작업으로 조사대상의 연령을 고려하여 설명분포함수의 중심분포을 균일분포(9,18)로 지정하고 설명분포를 구하기 위한 베타혼합체의 수를 $L = 5$로 설정하였다. 이 때 베타들의 가중값 ω는 한 값만 1이고 나머지는 0으로 만들어 다섯개의 베타분포 중 가장 적합력이 높은 베타분포를 선택하도록 만들었다. 연결분포함수의 중심분포 F는 균일분포(0,1)로 가정하고 $\psi(v,\lambda) = \mathrm{IB}(v;1,\lambda)$와 $\psi(v,\lambda) = \mathrm{IB}(v;\lambda,1)$를 추가적으로 이용하여 적합도를 높여보았다. 표 5.1는 R 프로그램을 이용하여 계산한 결과이다.

표 5.1: 변환 ψ에 대한 모수추정.

	모형선택		모수추정		Pearson의
ψ	$\hat{\beta}$	$\hat{\omega}$	$\hat{\lambda}$	$\hat{\sigma}$	X^2
Identity	1	3	.	2.17	161.21
IB(·,1)	1	2	2.65	1.62	15.43
IB(1,·)	1	3	1.57	1.98	23.52

로지스틱, 프로빗, 보로그로그 회귀모형에 대한 Pearson의 X^2 통계량의 값은 각각 21.31, 21.74, 190.93이다. 표 5.1에서 보는 것과 같이 IB(·,1) 변환을 적용한 모형에 대한 Pearson의 X^2 통계량이 15.43으로 기존의 표준 모형들 보다 자료 적합력이 높은 것으로 나타났다. 이것을 식으로 표시하면 다음과 같이 쓸 수 있다.

$$\mu = \begin{cases} 0, & \text{Age} < 9 \\ \left[\mathrm{IB}\left(\frac{\text{Age}-9}{9}, 3.24, 6.48\right)\right]^{2.65}, & 9 \le \text{Age} \le 18 \\ 1, & \text{Age} > 18. \end{cases}$$

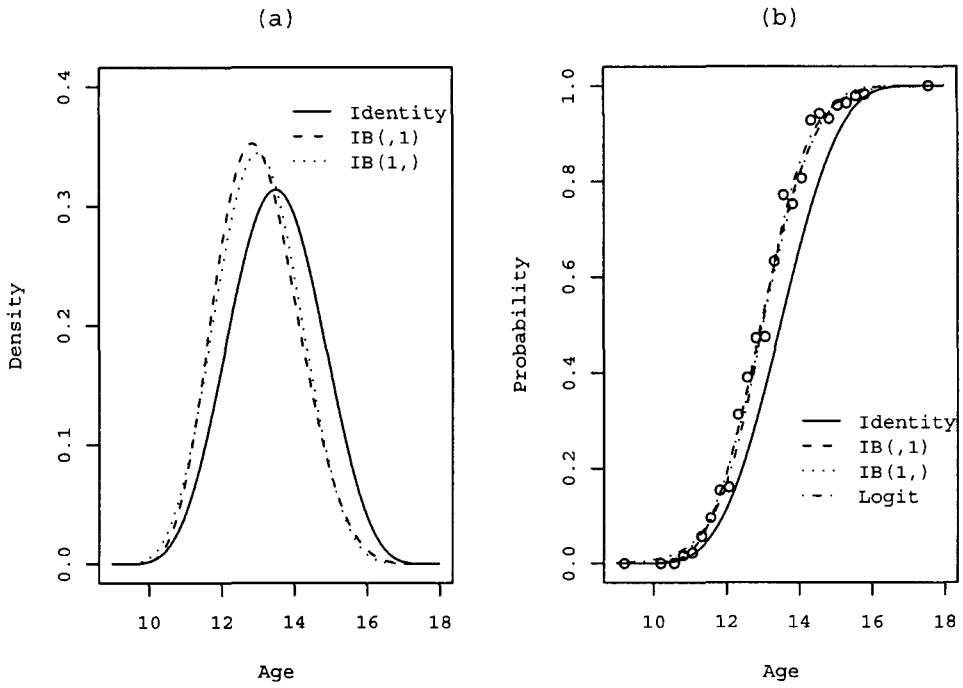

그림 5.1: ψ에 따른 일반화가중선형모형의 비교.

그림 5.1의 (a)는 μ를 설명변수 나이(Age)로 미분하여 얻은 밀도함수를 나타낸다. 이 그림에서 보는 것과 같이 IB(·,1)와 관련된 밀도함수는 12.84에서 최대값을 가지며 약간 양의 왜도를 가지는 형태를 가진다. 이것은 12.84세 이전에 초경이 발생할 확률은 나이가 많아질수록 계속 커지다가 12.84세 근처에서 가장 큰 확률을 가진 후 서서히 감소하는 형태를 가진다. Guerrero와 Johnson (1982)이 보인 것처럼 이 자료에 대해서는 대칭인 형태의 분포함수보다는 로그로그 척도와 같이 양의 왜도를 가지는 분포함수가 연결함수로 더 적합하다는 것을 의미한다. 로지스틱 회귀모형과 비교하여 B(·,1)의 모형에서의 Pearson X^2 통계량의 감소는 μ가 작은 부분에서 일어났다는 것을 그림 5.1의 (b)를 통해 알 수 있다. 추가적으로 bootstrapping pairs 방법을 이용하여 IB(·,1)의 모수 λ와 σ에 대한 추정값을 200개를 구한 결과 95% 붓스트랩 신뢰구간은 각각 [2.46,2.85]와 [1.48,1.75]로 나타났다.

6. 결론

이 논문에서는 반응평균을 분포함수로 변환시킨 값이 설명변수를 분포함수로 변환시킨 후 선형결합 형태로 가중한 것에 의해 관계가 설명되는 일반화가중선형모형을 소개하였다.

이 모형의 장점은 연결 관계가 이론적으로 문제를 발생시키지 않으며 각 설명변수의 설명 정도를 가중모수로 쉽게 해석할 수 있다는 것이다. 또한 베타혼합을 사용하여 자료로 하여금 설명분포함수를 선택하게 하여 반응평균에 대한 설명변수의 영향력이 어떻게 주어지는지에 대한 해석이 가능하다. 일반화선형모형에서 고려되고 있는 많은 문제들이 이 모형에서도 적용될 수 있다. 이 모형은 구조상 베이지안 방법을 적용하기 적합한 형태를 가지고 있는데 베이지안 관점에서 이 모형을 연구한 논문은 http://stat.chonbuk.ac.kr/inkwon에서 참고하기 바란다.

참고문헌

Yeo, I. K. (2001), Goodness of link tests for binary response data, <한국통계학회논문집>, **8**, 357-366.

Arando-Ordaz, F. J. (1981), On two families of transformations to additivity for binary response data, *Biometrika*, **68**, 357-363.

Cheng, K. F. and Wu, J. W. (1994), Testing goodness of fit for a parametric family of link function, *Journal of the American Statistical Association*, **89**, 657-664.

Guerrero, V. M. and Johnson, R. A. (1982), Use of the Box-Cox transformation with binary response models, *Biometrika*, **69**, 309-314.

John, J. A. and Draper, N. R. (1980), An alternative family of transformations, *Applied Statistics*, **29**, 190-197.

Lehmann, E. L. and Casella, G. (1998), *Theory of Point Estimation (revised edition)*, Springer-Verlag, New York.

Mallick, B. K. and Gelfand, A. E. (1994), Generalized linear models with unknown link functions, *Biometrika*, **81**, 237-245.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models* 2nd Ed. Chapman & Hall, New York.

Milicer, H. and Szczoka, F. (1966), Age at menarche in Warsaw girls in 1965, *Human Biology*, **38**, 199-203.

Nelder, J. A. and Wedderburn, R. W. M. (1972), Generalized linear models, *Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

Stukel, T. (1988), Generalized logistic models, *Journal of the American Statistical Association*, **83**, 426-431.

Taylor, J. (1988), The cost of generalized logistic regression, *Journal of the American Statistical Association*, **83**, 1078-1083.

Yeo, I. K. and Johnson, R. A. (2000), A new family of power transformations to improve normality or symmetry, *Biometrika*, **87**, 954-959.

[2003년 10월 접수, 2004년 4월 채택]

Generalized Weighted Linear Models Based on Distribution Functions - A Frequentist Perspective*

In-Kwon Yeo [1)]

ABSTRACT

In this paper, a new form of linear models referred to as generalized weighted linear models is proposed. The proposed models assume that the relationship between the response variable and explanatory variables can be modelled by a distribution function of the response mean and a weighted linear combination of distribution functions of covariates. This form addresses a structural problem of the link function in the generalized linear models in which the parameter space may not be consistent with the space derived from linear predictors. The maximum likelihood estimation with Lagrange's undetermined multipliers is used to estimate the parameters and resampling method is applied to compute confidence intervals and to test hypotheses.

Keywords: Bootstrap, Conjugate family, Exponential family, Mixture of beta distributions Parametric transformation.

* This work was supported by Korea Research Foundation Grant (KRF-2002-003-C00028)

1) Assistant Processor, Division of Mathematics and Statistics, Sookmyung Women's University, Chungpa Dong 2Ka 53-12, Yongsan-Ku, Seoul, 140-742, KOREA
E-mail: inkwon@sookmyung.ac.kr