Journal of Statistical
Theory & Methods
1997, Vol. 8, No. 2, pp. 195 ~ 209

평균과 산포의 동시 모형화에 대한 모형검토

하 일도 · 이 우동 · 조 건호 [1]

요약

일반화 선형모형의 범위를 크게 확장한 준-우도 모형에서 반응변수의 분산성분인 산포모수가 상수가아니라 어떤 공변량들의 값에 의존하여 변하는 경우, 평균과 산포의 동시 모형화가 요구된다. 본 논문에서는 준-우도 모형에서 평균과 산포의 동시 모형화를 통해 실제 자료를 쉽게 분석하도록 해주는 통계 패키지 GENSTAT(release 5.3.2, 1996)을 활용하여, Carrol과 Ruppert(1987,pp.46-47)에 의해 소개된 에스테르 분해효소 (esterase assay)의 자료에 대해 그래픽 방법을 이용한 모형검토를 통해서 기존의 평균모형 보다는 평균과 산포의 동시 모형화를 고려해야 하는 필요성을 언급한 뒤, 그 자료에 대한 적절한 평균과 산포의 동시 모형을 찾는 방법을 연구한다.

주제어: 분산함수, 일반화 선형모형, 준-우도, 평균과 산포의 동시 모형화, 편차잔차.

1. 서론

고전적 선형 모형(classical linear models)은 모평균에 대한 모형으로서 반응변수 Y가 다음 세 가지 조건을 만족한다고 가정한다.

(i) Y가 정규분포를 따른다. (ii) Y의 평균이 공변량들(covariates)의 선형함수로 표현된다. (iii) Y의 분산이 평균과 관계없는 상수인 산포모수(dispersion parameter)이다. 그러나 이 세 조건의 일부 또는 모두를 만족하는 자료는 실제 매우 드물다. Box와 Cox(1964)는 자료가 위의 세 조건의 일부 또는 모두를 만족시키지 않을 경우에는 이들을 충족시키도록 자료변환(data transformation)을 한 뒤 분석을 해야한다고 주장하였다. 하지만 Hougarrd(1982)가 밝혔듯이 이 세 가지 조건을 동시에 충족시킬 수 있는 자료변환은 존재하지 않는다.

이러한 어려움을 극복하기 위해 Nelder와 Wedderburn(1972)은 위의 세 조건을 각각 확장한 다음과 같은 일반화 선형모형(Generalized Linear Models; GLMs)을 개발하였다.

$(i)'$ Y가 자연 지수족(nature exponential famly)에 속하는 분포를 따른다. $(ii)'$ Y의 평균이 연관함수(link function)를 통해 공변량들의 선형함수로 표현된다. $(iii)'$ Y의 분산이 평

[1](712-240) 경북 경산시 점촌동 산 75 경산대학교 통계학과

균과 관계없는 상수인 산포모수와 Y의 평균에만 의존하는 분산함수 (variance function)의 곱이다.

GLMs의 개발로 인해, 정규분포를 따르는 연속형자료의 회귀분석 뿐만 아니라, 통상적으로 개수 (count)나 비율(ratio)자료가 각각 따르는 포아송과 이항분포 그리고 양의 값을 갖는 연속형 변수로 등변동계수(constant coefficient of variation)를 갖는 감마분포등을 이용하여 보다 다양한 자료를 회귀분석 할 수 있다.

나아가, Wedderburn(1974)은 GLMs의 분포적인 가정인 $(i)'$를 제외한, 반응변수의 평균과 분산만의 가정인 즉, $(ii)'$와 $(iii)'$만의 보다 완화된 가정을 바탕으로 한 준-우도 (Quasi-Likelihood; QL)함수를 제안함으로서 GLMs의 범위를 크게 넓혔다. 이러한 우도를 근거로 한 준-우도 모형(Quasi-Likelihood Models; QLMs)은 반응변수의 분포를 결정하기는 어렵지만 그것의 평균과 분산간의 관계가 정확하게 가정될 수 있는 자료의 분석에 폭 넓게 사용되어 진다.

한편, Nelder와 Pregibon(1987)은 QL이 똑같은 자료에 대해 서로 다른 산포모수 또는 분산함수를 갖는 모형의 공식적인 비교에 사용될 수 없다는 제약 때문에, 확장된 준-우도(Extended Quasi-Likelihood; EQL)함수를 소개하였다. Pregibon(1984)은 모든 반응값에 대해 산포모수가 상수인 가정을 갖는 QLMs에서 그 산포모수가 상수가 아니라 어떤 공변량들의 값에 의존하여 변할 수 도 있다는 사실로 인해 이러한 산포모수의 모형화 즉, QLMs에서 평균과 산포의 동시 모형화 를 처음으로 주장하였다. 특히, 고적적 선형모형에서는 반응변수의 분산은 분산함수가 1인 상수인 산포모수 즉 등분산(constant variance)이므로, 이때는 산포모형에 관한한 이분산(non-constant variance)에 대해 모형화 하는 경우가 된다. 그리고 EQL은 QLMs에서 이러한 동시 모형화의 추론(inference)에 대한 객관적인 기준(criterion)을 제공한다. 더욱이, Lee와 Nelder(1992)는 EQL을 사용하는것이 소표본의 자료분석에 좋다는 사실을 밝혔다. Aitkin(1987)은 고전적인 선형모형에서 평균과 산포의 동시 모형화(joint modelling of mean and dispersion)를 연구하였다. 특히, 제품간의 변이(variation)를 줄이고 제품의 평균을 성능 목표치에 맞추는 적정 공정을 찾아 내는데 그 목적이 있는 Taguchi(1985) 품질개선 실험에서 얻어진 자료 분석에 최근 Nelder와 Lee(1991), 이 영조(1993), 그리고 이 영조와 임 용빈(1996)등이 QLMs에서 이러한 동시 모형화를 적용하여 많은 응용 및 연구를 진행하고 있다.

통계적 자료분석에서 한 가지 유의할 점은 주어진 자료에 대해 모형검토 (model checking)를 거친 후 그 자료에 대한 적절한 분석 결론을 내려야 한다. 만약 이러한 모형검토를 거치지 않고 내린 분석 결론은 모순된 결론이 될 가능성이 아주 높기 때문이다. 일반적으로 모형검토는 잔차검토(residual checking)를 통해서 이루어진다.

본 논문에서는 준-우도 모형에서 평균과 산포의 동시 모형화를 통해 실제 자료를 쉽게 분석하도록 해주는 통계 패키지 GENSTAT(release 5.3.2,1996)을 활용하여, Carrol과 Ruppert(1987,pp.46-47)에 의해 소개된 에스테르 분해효소(esterase assay)의 자료에 대해 최종적으로 고려된 기존의 평균모형 보다는 평균과 산포의 동시 모형화를 고려해야 하는

필요성을 언급한 뒤, 그 자료에 대한 적절한 평균과 산포의 동시 모형을 찾는 방법을 연구한다. 2절에서는 QLMs에서 평균과 산포의 동시 모형화를 살펴보고, 3절에서는 모형검토를 용이하게 해주는 잔차그림(residual plot)을 언급하고, 다루고자 하는 예에 대한 자료 소개 및 GENSTAT을 활용하여 그래픽 방법을 이용한 모형검토를 통해서 그 자료의 적절한 평균과 산포의 동시 모형을 찾아가는 과정을 연구한다.

2. 평균과 산포의 동시 모형화

$Y_i(i=1,\cdots,n)$를 i 번째 반응변수라 하고 이들은 서로 독립이라 하자. 그러면 GLMs의 범위를 크게 확장한 QLMs은 다음과 같이 표현된다.

$$\eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j, \tag{1}$$

여기서, $E(Y_i)=\mu_i, Var(Y_i)=\phi V(\mu_i)$이고 η_i는 선형 예측식(linear predictor), $g(\cdot)$는 연관함수, x_{ij}는 i번째 관측값에 대한 j번째 공변량, β_j는 j 번째 회귀 모수이며, ϕ는 상수인 산포모수, $V(\cdot)$는 기지의 분산함수이다. 그리고 $q(\beta)$를 β_j들로 구성된 행벡터(row vector)인 β의 QL, 보다 정확히 준-로그-우도(quasi-log-likelihood), 라 표기하면 Wedderburn(1974)의 QL 방정식은 다음과 같이 주어진다.

$$\partial q(\beta)/\partial\beta_j = \sum_{i=1}^{n}\{(y_i-\mu_i)/V(\mu_i)\}(\partial\mu_i/\partial\beta_j) = 0 \tag{2}$$

방정식 (2)의 해(solution)인 최대 준-우도추정량(maximum quasi-likelihood estimator;MQLE)은 반복적인 가중 최소 자승 절차(iterative weighted least squares procedure)에 의해 얻어지며, 식 (2)는 GLMs과 똑같은 추정방정식(estimating equation)이므로 MQLE는 GLMs에서의 최대 우도 추정량(maximum likelihood estimator)과 같다. McCullagh(1983)은 적절한 조건하에서 MQLE의 점근 정규성(asymtotic normality)을 밝혔다. 특히, QL은 $Var(Y_i)=\phi V(\mu_i)$를 갖는 GLMs형태의 분포가 존재한다면 진짜 우도(true likelihood)가 될 수 있다. QLMs의 보다 자세한 성질은 Wedderburn(1974), McCullagh(1983) 그리고 McCullagh와 Nelder(1989, chap. 9)등을 참조하기 바란다.

한편, Pregibon(1984), Nelder와 Pregibon(1987) 그리고 McCullagh와 Nelder (19- 89, chap.10)는 QLMs (1)에서 산포모수 ϕ가 더 이상 상수가 아니라 어떤 공변량에 의존하여 변한다는 가정에 의해 이를 모형화한 다음과 같은 QLMs의 평균, 산포 동시 모형화를 고려하였다.

$$\eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j, \quad \zeta_i = h(\phi_i) = \sum_j u_{ij}\gamma_j, \tag{3}$$

여기서, $E(Y_i) = \mu_i, Var(Y_i) = \phi_i V(\mu_i)$이고, ζ와 $h(\cdot)$는 각각 산포 예측식과 산포 연관함수이며, u_{ij}는 모수 γ_j를 갖는 산포 공변량이며 (3)의 왼쪽항에서의 x_{ij}들의 부분집합이 될 수도 있다. 즉, 산포공변량은 평균 공변량과 같거나 다를 수 도 있다. 참고적으로 모형 (3)의 왼쪽(오른쪽)항을 평균(산포)모형이라고 각각 부른다. 산포 연관함수로서 일반적으로 로그를 택하는 이유는 산포모수가 항상 양수임을 보장해 주기 때문이다. 물론 항등(identity) 또는 역수(reciprocal) 연관함수등도 가능은 하다. 모형 (3)의 추론에 객관적인 기준을 제공하는 Nelder와 Pregibon(1987)의 EQL(q^+)은 다음과 같이 정의된다.

$$-2q^+ = \sum_{i=1}^{n} \{(d_i/\phi_i) + log(2\pi\phi_i V(y_i))\} \tag{4}$$

여기서, d_i는 평균모형에서의 i번째 편차 요소(deviance component) 즉, $d_i = 2\{q(y_i; y_i) - q(\mu_i; y_i)\}$. EQL (4)로 부터, β_j에 대한 추정방정식을 구하면

$$\partial(2q^+)/\partial\beta_j = \sum_{i=1}^{n} \{(y_i - \mu_i)/(\phi_i V(\mu_i))\}(\partial\mu_i/\partial\beta_j) = 0, \tag{5}$$

으로서 식 (5)는 $1/\phi_i$이 가중치(weigh)로 포함된다는 사실을 제외하고는 Wedderburn의 QL 방정식 (2)와 같으며, γ_j에 대한 추정방정식은 다음과 같이 구해진다.

$$\partial(2q^+)/\partial\gamma_j = \sum_{i=1}^{n} \{(d_i - \phi_i)/\phi_i^2\}(\partial\phi_i/\partial\gamma_j) = 0. \tag{6}$$

식 (6)은 반응변수로서 d_i를 갖는 $V(\mu_i) = \mu_i^2$ 에 대한 Wedderburn의 QL 방정식이 되며, 이 식의 해는 반응변수 d_i가 감마분포 즉 $\phi_i\chi_1^2$분포를 따른다고 가정함으로서 얻어질 수 있다. 사실, d_i가 근사적으로 이러한 분포를 따르는 경우가 많으며 특히, 반응변수가 정규분포를 따르는 경우 d_i는 정확하게 이러한 감마분포를 가진다. 그리고, 이 식 역시, 반복적인 가중 최소 자승 절차를 사용하여 해를 얻을 수 있다. 결국, (5)와 (6)의 추정 절차는 먼저, 주어진 사전 가중치 $1/\hat{\phi}_i$에 대해 (5)의 해를 구하고, 다음으로 반응변수 $d_i = d_i(Y_i, \hat{\mu}_i)$를 사용하여 (6)의 해를 구한 뒤, (6)에서 얻어진 해가 다시 (5)의 사전 가중치(prior weight)로 주어져서 새로운 (5)의 해를 얻는 반복적인 방법(iterative method)이 요구된다. 하지만 우리는 GENSTAT을 사용하여 적합된 평균과 분산 모형을 쉽게 얻을 수 있다. 마지막으로, QLMs의 평균 산포 동시 모형을 요약하면 표 1과 같다.

덧붙여, 이러한 평균 산포 동시 모형화에 대한 보다 자세한 내용은 Nelder와 Pregibon(1987), McCullagh와 Nelder(1989, chap. 10), Nelder와 Lee(1991)등을 참조하기 바란다.

표 1: QLMs의 평균 산포 동시 모형

구성요소	평균모형	산포모형
반응변수	Y_i	d_i
평균값	μ_i	ϕ_i
산포값	ϕ_i	2
분산함수	(임의)	ϕ_i^2
연관함수	(임의)	log(일반적)
선형 예측식	$\eta_i = \sum_j x_{ij}\beta_j$	$\zeta_i = \sum_j u_{ij}\gamma_j$
사전 가중치	$1/\phi_i$	1

3. 모형검토

3.1 잔차그림

일반적으로 모형검토는 잔차그림의 검토를 통해 쉽게 파악할 수 있으며, GLMs형태의 모형에서 주로 사용하는 잔차는 피어슨잔차(pearson residual)와 편차잔차(deviance residual)이지만 후자를 사용하는 것이 더 좋다; Pierce와 Schafer (1986)는 후자가 정규분포에 더 가깝다는 사실을 보였을 뿐만 아니라 우도(likelihood)에 근거한 방법에 대해 후자는 자연스러운 선택이 된다고 지적하였다. 따라서 편차잔차에 근거한 다음과 같은 유용한 잔차그림의 검토를 통해 모형검토를 할 수 있다:

(a): 연관함수 또는 선형 예측식은 $\hat{\eta}$ 또는 적합값의 함수인 척도적합값(scaled fitted value)에 대한 표준화된 편차잔차(standardized deviance residual)의 그림으로 검토될 수 있다. 이 그림에서 곡선(curvature)과 같은 특별한 추세(trend)가 없다면 만족된 연관함수 또는 선형 예측식이라고 볼 수 있다. 그리고 평활 (smoothing)은 그 추세의 파악에 유용할 것이다. (b): 분산함수의 검토는 척도적합값에 대한 표준화된 편차잔차의 절대값의 그림으로 파악할 수 있다. 만족된 분산함수는 평균에 대해 특별한 추세를 보이지 않지만 잘못된 분산함수는 어떤 추세로 나타날 것이다. 양의 추세(positive trend)는 현재의 분산함수가 평균에 대해 다소 느리게 증가하고 있음을 지시한다; 예를들어, $V(\mu) \propto \mu$의 본래의 선택(original choice)은 $V(\mu) \propto \mu^2$으로 대치될 필요가 있을 것이다. 물론, 음의 추세(negative trend)는 그 반대를 지시할 것이다. (a)에서와 같이 평활은 이러한 추세를 보다 명확하게 파악하는데 도움을 줄 것이다. (c): 표준화된 편차잔차의 정규성은 반-정규(half-Normal) 또는 완전 정규그림(full Normal plot)으로 검토할 수 있으며, 그 잔차들이 거의 직선의 형태를 띠면 그것의 정규성이 만족된다고 할 수 있다. 특히, 반-정규그림은 이상치(outlier)판별에 보다 유용하게 사용된다.

위와 같은 세 가지 잔차검토에서 특별한 이상이 없을 때, 주어진 자료에 대해 적합된 모

형은 적절하다고 할 수 있을 것이다.

위의 (a)-(c)의 그림은 GENSTAT에서 제공해 주며, 모형검토에 대한 보다 자세한 내용은 McCullagh와 Nelder(1989, chap. 12)를 참조하기 바란다.

우리는 하나의 예를 갖고서 위와 같은 모형검토를 통해 적절한 모형을 찾아보려고 한다.

3.2 에스테르 분해효소 시금의 예

Carrol과 Ruppert(1987, pp.46-47)에 의해 소개된 에스테르 분해효소 시금 (esterase assay)의 예는 에스테르 분해효소의 농도(x: 설명변수)를 관측한 뒤 한 결합실험(binding experiment)으로 부터 그 효소의 농도에 대한 결합 개수(y: 반응변수)가 얻어진 자료(108개)에 대해, x와 y의 적절한 모형을 찾아 이를 이용하여 주어진 $y(y_0)$에 대응하는 $x(x_0)$를 추론(inference) 하는데 그 목적이 있는 보정(calibration)문제이다. 우리는 여기서 보정문제보다는 x와 y의 적절한 모형를 어떻게 찾을 것인가에 대해 궁극적인 목적이 있다. Carrol과 Ruppert는 항등 연관함수를 갖는 등변동계수 모형, 즉 $\mu = E(Y) = \beta_0 + \beta_1 x, Var(Y) = \phi\mu^2$을 적절한 모형으로 택하여 보정문제를 다루었다. 우리는 이러한 등변동계수 모형을 검토하기 위해, GENSTAT을 이용하면 그림 1과 같은 잔차그림을 얻을 수 있다. 그림 1의 왼쪽(오른쪽)상단은 각각 (a)((b))를 각각 검토해 주며, 왼쪽 및 오른쪽 하단은 각각 반 및 완전 정규그림을 가르키며 (c)를 파악하게 해 준다. 그림 1에 의하면, 오른쪽 상단으로 부터 평균 증가에 따라 분산함수가 어느 정도 증가함을 알 수 있고, 완전 정규그림으로 부터 직선에서 벗어남을 파악할 수 있으며 또한 정규성을 검정하는 샤피로-윌크(Shapiro-Wilk)통계량의 값이 0.9533(p-값=0.0034)이므로 표준화된 편차잔차가 정규분포를 따르지 않음을 알 수 있다. 따라서 우리는 예1의 자료에 대해 다음과 같은 준-우도 모형들을 적합시켜서 적절한 모형을 찾아 보려고 한다.

$$\text{모형}1 : \mu = E(Y) = \exp(\beta_0 + \beta_1 x), \quad Var(Y) = \mu.$$

반응값이 개수이므로 로그 연관함수를 갖는 포아송모형인 모형 1을 적합시켜 본 결과 편차(deviance) 4057에 자유도(degree of freedom) 106으로 과다산포(over-dispersion)가 일어나므로 이 모형이 적절치 않음을 알 수 있다. 또한 그림 2에 의하면 표준화된 편차잔차의 값이 거의 대부분 크며, 왼쪽 상단 그림으로 부터 이차항의 설명변수 또는 다른 연관함수가 요구됨을 알 수 있고 오른쪽 상단으로 부터 분산함수의 지정이 잘못 되었음을 알 수 있다. 하나의 대안으로서 과산포된(over-dispersed)포아송 모형 즉 $\mu = E(Y) = \exp(\beta_0 + \beta_1 x), Var(Y) = \phi\mu$를 고려할 경우 그 잔차값의 크기만 줄어들뿐 그림 2의 형태가 그대로 유지된다. 이것은 포아송 모형과 과산포된 포아송 모형에서 회귀계수의 추정량은 똑같고 그것의 표준오차(standard error)와 표준화된 편차잔차만 $\hat{\phi}$ = 편차/자유도 = 38.27만큼 과산포된 포아송

에서 줄어들기 때문이다.

$$\text{모형}2 : \mu = E(\log(Y)) = \beta_0 + \beta_1 \log(x), \quad Var(\log(Y)) = \phi.$$

그림 3의 x에 대한 y의 산점도로 부터 선형이기 보다는 이분산이 명확하게 보이므로, 두 변수 모두 로그 변환을 한 결과 어느정도 선형 형태를 띠고 있음을 그림 4로 부터 알 수 있다. 따라서 고전적 로그 선형모형인 모형 2를 적합시켜 그림 5의 잔차그림을 얻었지만 여전히 이분산이 어느정도 보인다. 또한, 그림 5는 그림 1과 비슷한 결과를 준다. 그 이유는 $Var(Y) \simeq \phi\mu^2$이면 $Var(log(Y)) \simeq \phi$이기 때문이다. 즉, 모수변환을 통한 등변동계수 모형과 자료변환을 통한 고전적 로그 선형모형은 거의 비슷한 자료분석의 결과를 제공한다.

$$\text{모형}3; \mu = E(Y) = \exp(\beta_0 + \beta_1 \log(x)), \quad Var(Y) = \phi.$$

로그 연관함수를 갖는 고전적 선형모형인 모형 3을 적합시켜 그림 6을 살펴 본 결과 다음과 같은 사실을 알 수 있다. 즉, 그림 6의 왼쪽 상단 그림에서 평활선이 어느 정도 수평 추세를 보이므로 모형 3의 왼쪽항인 모평균은 적절하지만 그림 6의 오른쪽 상단 그림에서의 평활선이 증가 추세를 보이므로 모형 3의 오른쪽항인 반응변수의 분산은 등분산이 아닌 이분산을 지시한다. 부가적으로 그림 6의 반 및 완전 정규그림으로 부터 3개의 이상치가 존재함을 파악 할 수 있다. 따라서 모형 3의 모평균은 그대로 유지하고 반응변수의 분산 즉 산포를 모형화 할 필요성이 요구되므로 다음과 같은 평균과 산포의 동시모형을 고려해 본다.

$$\text{모형}4 : \log\mu = \beta_0 + \beta_1 \log(x), \ \log\phi = \gamma_0 + \gamma_1 \log(x) \ (\text{단}, E(Y) = \mu, Var(Y) = \phi)$$

모형 4의 평균 및 산포모형에 대한 잔차그림은 각각 그림 7.1과 7.2로 주어진다. 이러한 두 그림 으로 부터 특별한 문제점이 발견되지 않음을 알 수 있으므로 모형 4가 예 1의 자료에 적절한 모형이 될 수 있겠다. 모형 4를 적합한 결과는 다음과 같으며, 괄호 속의 값은 각 회귀계수의 추정량에 대한 표준오차(standard error)이고 그 회귀계수들 모두 매우 유의함을 알 수 있다:

$$\log\hat{\mu} = 2.340 + 1.144\log(x), \quad \log\hat{\phi} = 2.035 + 2.325\log(x)$$
$$(0.154) \quad (0.053) \qquad\qquad (0.789) \quad (0.271)$$

덧붙여, 또다른 하나의 적절한 모형으로서 다음과 같은 모형 5를 고려할 수 있는데 이는 모형 1의 잔차그림인 그림 2의 왼쪽 상단 그림에서 이차항의 설명변수를 요구하고 있다고 생각했기 때문이다.

$$\text{모형}5 : \log\mu = \beta_0 + \beta_1 x + \beta_2 x^2, \ \log\phi = \gamma_0 + \gamma_1 x \ (\text{단}, E(Y) = Var(Y) = \mu)$$

모형 5를 적합시켜 본 결과 그림 8.1과 8.2로 부터 모형 4보다 다소 더 적절한 것 같다. 따라서 모형 5를 적합시킨 다음의 결과로 부터 회귀계수들은 모두 매우 유의함을 알 수 있다:

$$\log\hat{\mu} = 3.848 + 0.126x - 0.001x^2, \quad \log\hat{\phi} = 2.225 + 0.047x.$$
$$(0.137)\ (0.012)\ (0.0002) \qquad (0.291)\ (0.013)$$

결국, 모형 4와 5로 부터 설명변수 x인 에스테르 분해효소의 농도는 평균과 산포에 동시적으로 영향을 줌을 알 수 있으며, 참고적으로 이를 회귀분석 측면에서 살펴보면 x의 값이 클수록 y에 대한 개수오차가 커지게 되므로 x의 증가에 따른 분산의 증가는 당연하다고 할 수 있을 것이다.

4. 결론

GENSTAT에서 제공하는 아주 유용한 잔차그림을 통해, QLMs에서의 평균과 산포의 동시모형화에 대한 모형검토를 거쳐 적절한 모형을 찾을 수 있음을 하나의 간단한 예를 갖고서 살펴보았다. 3.2절에서 설명변수가 하나인 경우에 평균과 산포의 동시 모형화를 고려하였지만 그 변수가 둘 이상인 경우에도 이러한 동시모형을 바로 적용할 수 있다. 특히, 기존의 평균모형으로 실제 자료를 적절하게 적합시키지 못하는 경우 이러한 동시 모형을 고려 할 수 있을 것이며, 이 모형을 통해 보다 다양한 자료를 분석할 수 있을 것으로 기대된다.

참고문헌

1. 이 영조 (1993). 다구찌 실험분석에 있어서 일반화 선형모형 대 자료변환,응용통계연구, 6권, 2호, 253-263.

2. 이 영조, 임 용빈 (1996). 일반화 선형모형을 통한 품질개선 실험 자료분석,품질경영학회지, 24권, 2호, 128-141.

3. Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM, *Applied Staticts,* 36, 332-339.

4. Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations(with discussions), *Journal of the Royal Statistical Society B,* 26, 211-243.

5. Carroll, R. J. and Ruppert, D. (1987). *Transformation and weighting in regression,* London: Chapman and Hall.

6. Hougaard, P. (1982). Parameterizations of non-linear models, *Journal of the Royal Statistical Society B,* 44, 244-252.

7. McCullagh, P. (1983). Quasi-likelihood functions, *Annals of Statistics,* 11, 59-67.

8. McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models,* 2nd edn., London: Chapman and Hall.

9. 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.

10. Nelder, J. A. and Lee, Y. (1992), Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons, *Journal of the Royal Statistical Society B,* 54, 273-284.

11. Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika,* 74, 221-232.

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

13. Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association,* 81, 977-986.

14. Pregibon, D. (1984). Review of generalized linear models, *Annals of Statistics,* 12, 1589-1596.

15. Taguchi, G. (1985). Quality engineering in Japan, *Communications in Statistics,* A, 14, 2785-2801.

16. Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method, *Biometrika,* 61, 439-447.

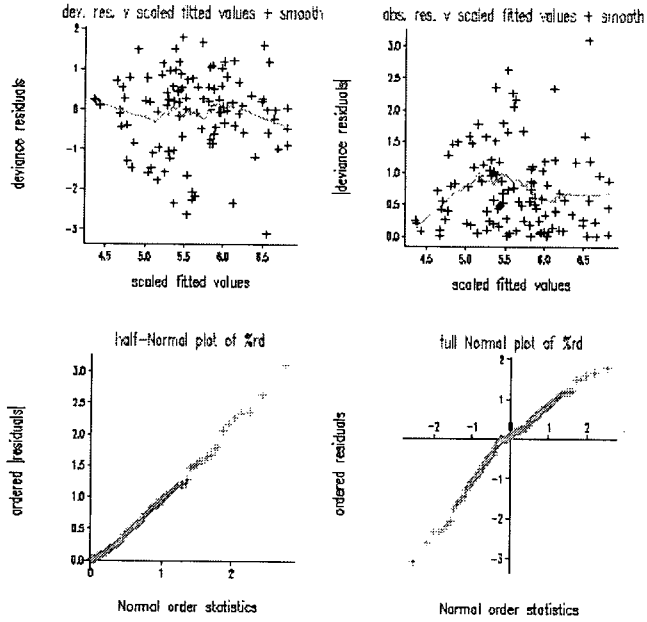

그림 1. 항등 연관함수를 갖는 등변동계수 모형
$(\mu = E(Y) = \beta_0 + \beta_1 x, Var(Y) = \phi\mu^2)$의 잔차그림

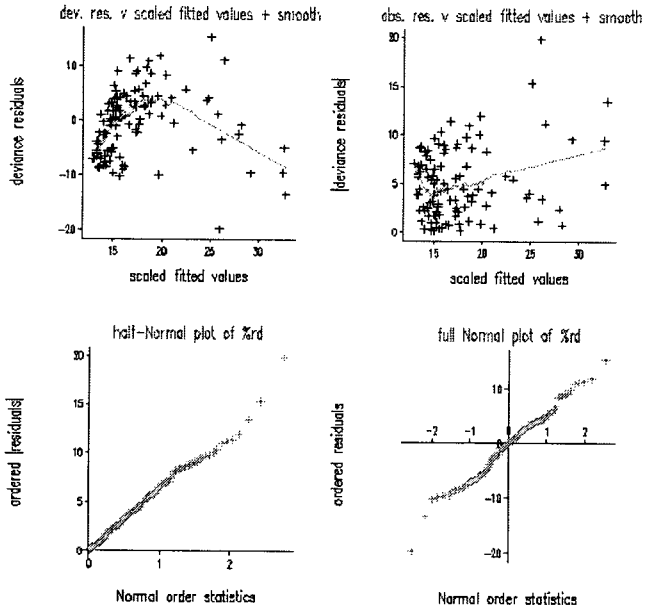

그림 2. 모형 1의 잔차그림

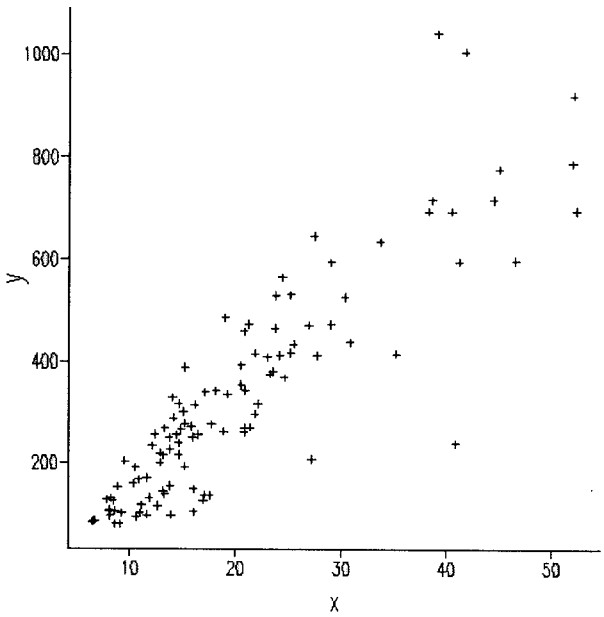

그림 3. x와 y의 산점도

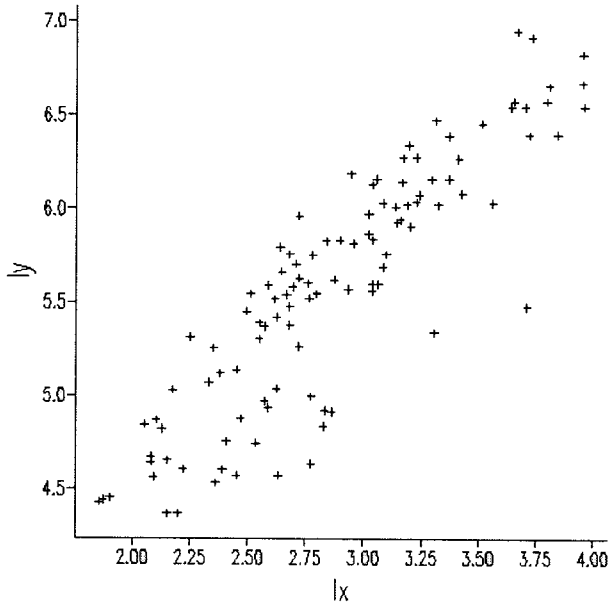

그림 4. $\log(x)(= lx)$와 $\log(y)(= ly)$의 산점도

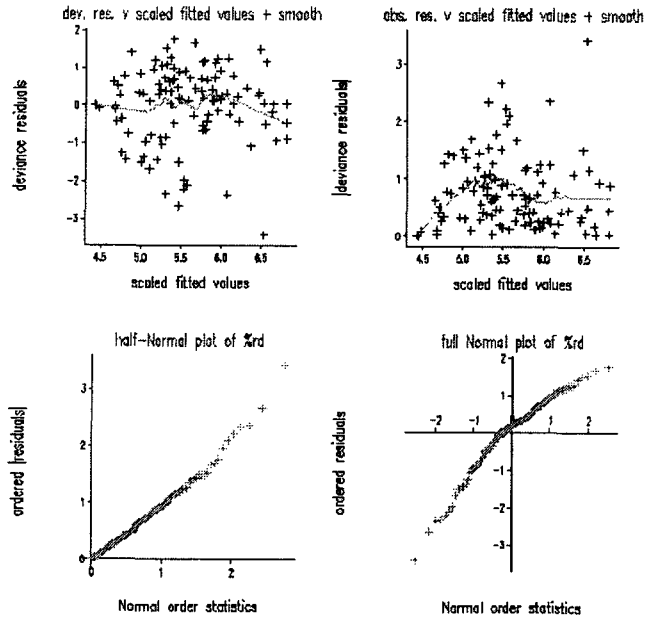

그림 5. 모형 2의 잔차그림

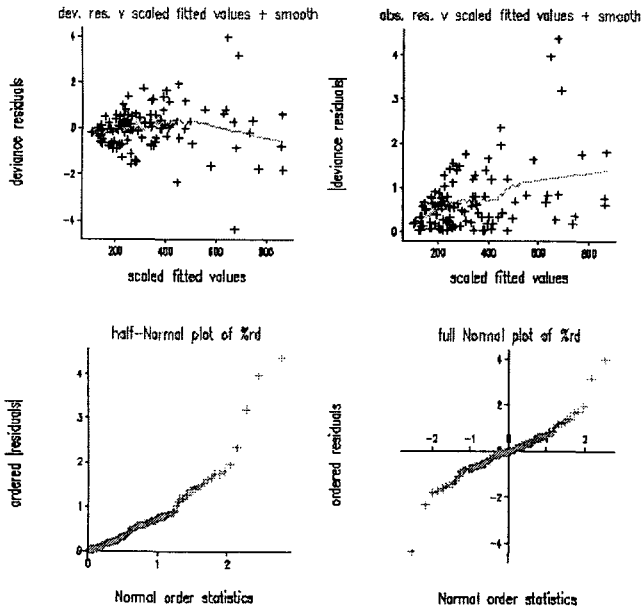

그림 6. 모형 3의 잔차그림

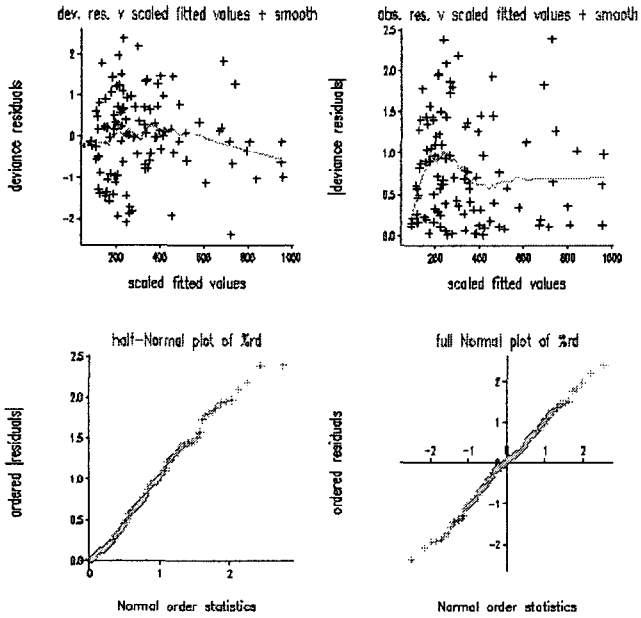

그림 7.1. 모형 4의 평균모형에 대한 잔차그림

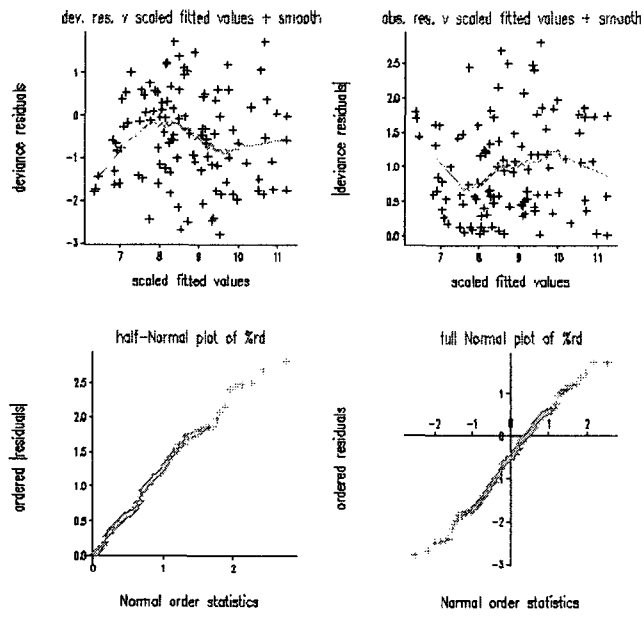

그림 7.2. 모형 4의 산포모형에 대한 잔차그림

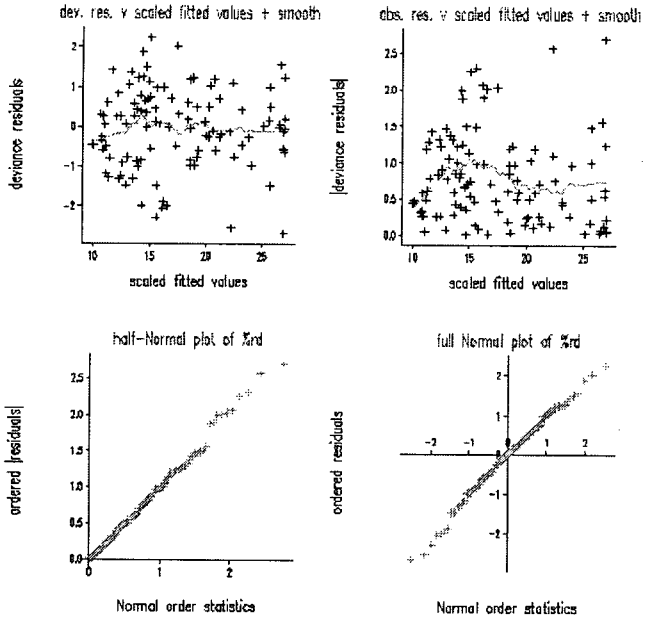

그림 8.1. 모형 5의 평균모형에 대한 잔차그림

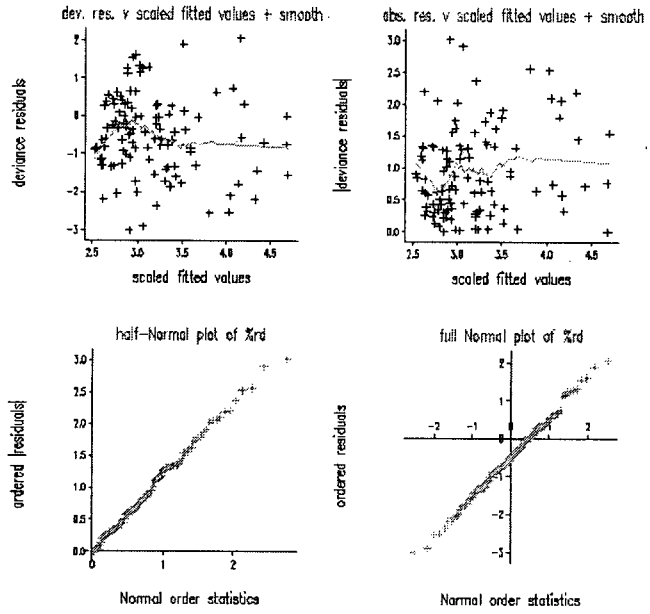

그림 8.2. 모형 5의 산포모형에 대한 잔차그림

Model Checking for Joint Modelling of Mean and Dispersion

Il-Do Ha · Woo-Dong Lee · Geon-Ho Cho [2]

Abstract

The joint modelling of mean and dispersion in quasi-likelihood models which greatly extend the scope of generalized linear models, is required in case that the dispersion parameter, the variance component of response variables, is not constant but changes by depending on any covariates. In this paper, by using statistical package GENSTAT(release 5.3.2, 1996) which makes a easily analyze real data through this joint modelling, we mention necessities that must consider this joint modelling rather than existing mean models through model checking based on graphic methods for esterase assay data introduced by Carrol and Ruppert(1987, pp.46-47), and then study methods finding reasonable joint model of mean and dispersion for this data.

[2]Department of Statistics, Kyungsan University,Kyungpook,712-240,Korea.