응용통계연구 제 13권 2호, 2000년, pp. 407-414

평균과 분산의 동시모형에 따른 회귀진단법에 관한 연구 *

강위창 [1] 이영조 [2] 송문섭 [3]

요 약

Carroll과 Ruppert(1988)는 준가능도(quasi-likelihood)를 이용하여 에스트라제 측정자료를 회귀분석하였다. Jung과 Lee(1997)는 준가능도을 이용한 회귀분석모형의 적합도검정통계량을 제안하였으며 검정 결과 기각되지 않아 본 분석모형이 타당하다고 주장하였다. 그러나 Lee와 Nelder(1998)의 잔차그림을 검토한 결과, 상기 모형으로는 평균증가에 따른 분산증가를 충분히 반영할 수 없었다. 본 논문에서는 Lee와 Nelder(1998)의 평균과 분산의 동시모형으로 에스트라제 자료를 재분석하고 잔차그림을 이용하여 모형의 타당성을 평가하였다. 또한 분산에서 산포모형에 대한 적합도검정에는 Lee와 Nelder(1998)의 제한가능도(restricted likelihood)에 근거한 검정법이 보다 적절함을 예시하였다.

주요용어: 수정단면 확장된 준가능도, 평균과 분산의 동시모형, 확장된 준가능도.

1. 서론

혈중 에스트라제 농도의 계측 자료를 분석하기 위하여, Carroll과 Ruppert(1988)는 평균과 분산함수에 대한 가정만으로 준가능도를 구성하여, 혈중 에스트라제 농도에 따른 단백질 결합수(binding count)를 회귀모형으로 분석하였다. 그림 1.1는 혈중 에스트라제 농도와 해당 결합수의 산점도이다. 에스트라제 농도 증가에 따라 결합수와 분산이 동시에 증가하므로 이를 설명하기 위하여 Carroll과 Ruppert(1988)는 다음 모형 (1.1)을 고려하였다.

$$E(y_i) = \mu_i = \beta_0 + \beta_1 x_i, \quad Var(y_i) = \phi V(\mu_i) = \phi \mu_i^2, \tag{1.1}$$

단, x_i는 에스트라제 농도, y_i는 해당 농도에서의 결합수, ϕ는 산포모수이며, $V()$는 분산함수이다. 모형 (1.1)에서는 그림 1.1의 평균증가에 따른 분산증가를 분산함수 $V(\mu_i) = \mu_i^2$을 통하여 모형화 하였다. Jung과 Lee(1997)는 모형 (1.1)에서 평균과 분산함수의 적합도를 동시에 검정하기 위한 통계량을 제안하였고, 이를 이용하여 적합도검정을 실시한 결과 기각되지 않아 모형 (1.1)이 타당하다고 주장하였다.

Nelder와 Lee(1991, 1997), Lee와 Nelder(1998)는 모형 (1.1)을 포함하는 일반적인 평균과 분산의 동시모형(joint modelling of mean and dispersion)을 제안하였고, 데비언스 잔

* 이 연구는 1998년도 교육부 기초과학 육성 연구비(1998-015-D00046)지원에 의한 것임.

1) (138-736) 서울시 송파구 풍납동 388-1, 서울중앙병원 의학통계연구실
E-mail: weechang@www.amc.seoul.kr
2) (151-742) 서울시 관악구 신림동 산 56-1, 서울대학교 자연과학대학 통계학과, 부교수
E-mail: youngjo@plaza.snu.ac.kr
3) (151-742) 서울시 관악구 신림동 산 56-1, 서울대학교 자연과학대학 통계학과, 교수
E-mail: songms@plaza.snu.ac.kr

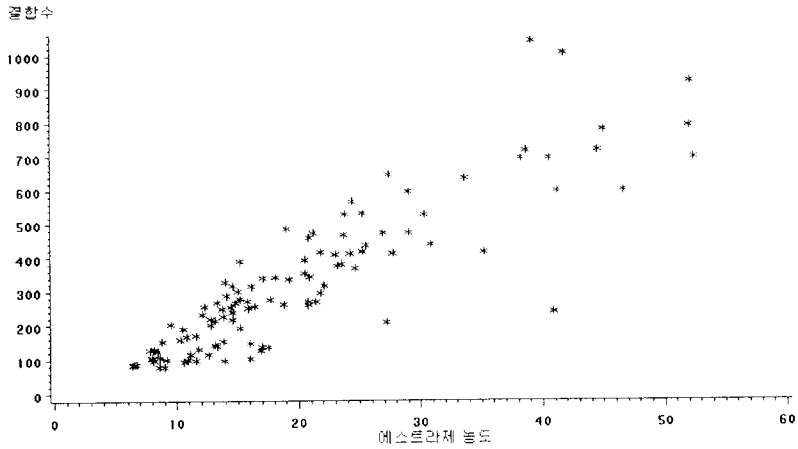

그림 1.1: 혈중 에스트라제 농도에 따른 결합수의 산점도

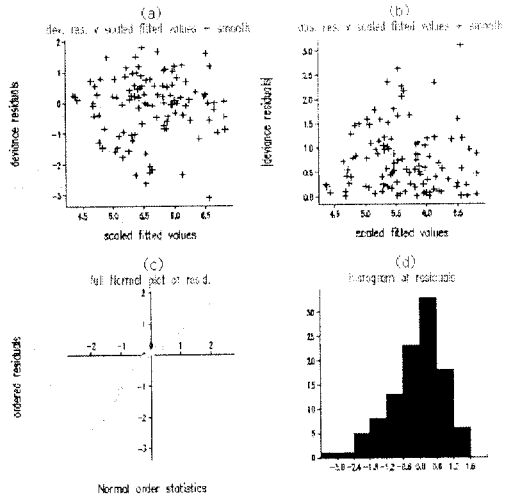

그림 1.2: 모형 (1.1)에 대한 잔차그림

차(deviance residual)에 근거한 표준화된 잔차(standardized residual)를 이용하여 모형검토를 위한 잔차그림들을 개발하였다. 이 표준화된 잔차는 McCullach와 Nelder(1989)의 식 (12.5)에 정의되어 있다. 그림 1.2는 모형 (1.1)에 대한 Lee와 Nelder(1998)의 모형검토를 위한 잔차그림들이다. 모형 (1.1)이 에스트라제 자료의 분석모형으로 타당하다면, 표준화된 잔차는 등분산성을 가져야 한다. 그러나 그림 1.2에서 적합값 $\hat{\mu}_i$에 대한 표준화된 잔차의 그림인 (a)를 보면, 잔차의 분산이 적합값의 증가에 따라 증가하였다. 또한 적합값에 대한 표

준화된 잔차의 절대값 그림인 (b)에서도, 적합값의 증가에 따라 잔차 절대값의 평균이 증가하는 경향이 나타나 (a)에서의 잔차의 분산증가를 뚜렷이 반영한다. 이는 그림 1.1에 나타난 평균증가에 따른 분산증가를 Carroll과 Ruppert(1988)의 분산모형 $Var(y_i) = \phi\mu_i^2$ 만으로는 충분히 고려할 수 없다는 것을 의미하며, 또한 Jung과 Lee(1997)의 검정통계량이 모형 (1.1)에 대한 적합도검정을 민감하게 수행하지 못했음을 반영한다고 생각한다.

본 논문에서는 그림 1.2에 나타난 것과 같은 반영되지 못한 분산증가를 모형화 할 수 있는 Lee와 Nelder(1998)의 평균과 분산의 동시모형으로 에스트라제 자료을 재분석하고, 잔차그림을 이용한 평균과 분산모형의 검토를 실시하여 선택된 모형이 보다 타당함을 보인다. 또한 분산모형의 검토에는 Lee와 Nelder(1998)의 제한가능도에 근거한 검정법이 보다 적절함을 예시한다.

2. 평균과 분산의 동시모형을 이용한 에스트라제 자료분석 및 모형검토

에스트라제 자료에서 반응변수는 갯수로 관측되므로, 그림 1.1의 평균에 따른 분산증가를 포아송 분산함수 $V(\mu_i) = \mu_i$에 근거한 과대산포(over-dispersion)모형 $Var(y_i) = \phi_i\mu_i$으로 설명하고자 한다. 그리고 μ_i는 양수이므로 로그 연관함수를 사용하는 것이 좋을 것이다. 이 경우 μ_i의 로그값은 설명변수 로그값의 일차함수로 표현될 수 있으므로 다음과 같은 평균과 분산의 동시모형을 생각하였다.

$$\text{평균모형} : \eta_i = \log(\mu_i) = \beta_0 + \beta_1 \log(x_i), \tag{2.1}$$

$$\text{산포모형} : \zeta_i = \log(\phi_i) = \gamma_0 + \gamma_1 \log(x_i). \tag{2.2}$$

모형 (1.1)에서는 산포모수 ϕ_i를 에스트라제 농도에 상관없이 일정한 상수 ϕ로 가정하였으나, 산포모형 (2.2)는 ϕ_i가 에스트라제 농도의 증가에 따라 변화한다고 가정하였다. 그러므로 그림 1.2에서와 같은 적합값 $\hat{\mu}_i$의 증가에 따른 표준화된 잔차의 반영되지 못한 분산증가를 산포모형 (2.2)를 통하여 설명할 수 있겠다.

평균모형 (2.1)과 산포모형 (2.2)의 추론은 각각 확장된 준가능도(extended quasi-likelihood)와 수정단면 확장된 준가능도(adjusted profile extended quasi-likelihood)를 이용하여 수행한다(Lee와 Nelder, 1998, 2장). 표 2.1는 두 모형 (2.1)과 (2.2)로 에스트라제 자료를 분석한 결과이다. 표 2.1로부터 에스트라제 농도의 로그값이 평균 결합수와 산포모수에 유의한

표 2.1: 평균과 분산의 동시모형 (2.1)과 (2.2)를 이용한 에스트라제 자료분석 결과

평 균 모 형				산 포 모 형			
요인	$\hat{\beta}$	표준오차	t 값	요인	$\hat{\gamma}$	표준오차	t 값
Constant	2.334	0.156	14.82	Constant	-0.364	0.789	-0.462
$\log(x_i)$	1.146	0.054	21.09	$\log(x_i)$	1.218	0.271	4.497

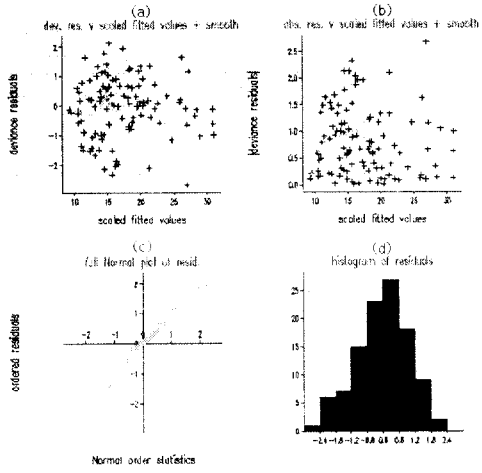

그림 2.1: 평균모형 (2.1)에 대한 잔차그림

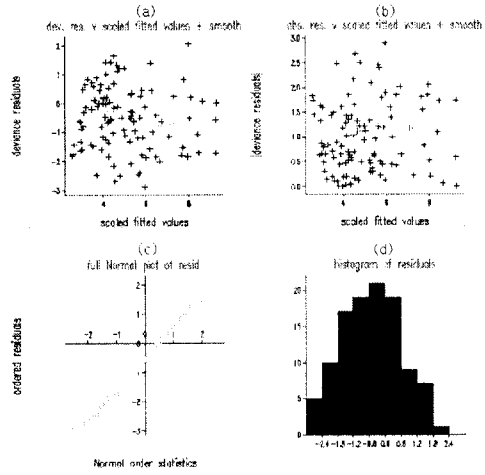

그림 2.2: 산포모형 (2.2)에 대한 잔차그림

영향을 준다는 것을 알 수 있고, 또한 산포모수 γ_1의 추정치가 양수라는 것으로부터 에스트라제의 농도가 증가함에 따라 분산함수 $V(\mu_i) = \mu_i$를 통한 분산증가 이외의 유의한 추가적인 분산증가가 있음을 알 수 있다.

그림 2.1은 평균모형 (2.1)에 대한 잔차그림들이다. 적합값에 대한 표준화된 잔차그림 2.1 (a)를 보면, 모형 (1.1)의 잔차그림 1.2 (a)에서 나타났던 적합값의 증가에 따른 잔차분산의 증가는 뚜렷하게 감소하였음을 알 수 있다. 또한 적합값에 대한 표준화된 잔차의 절대

값 그림 2.1 (b)에서는 그림 1.2 (b)에 나타났던 잔차 절대값의 평균이 증가하는 현상이 거의 사라졌으며, 이는 적합값의 증가에 따른 잔차분산의 증가가 더 이상 나타나지 않는다는 것을 반영한다. 즉, 그림 1.2 (a), (b)에서 나타났던 적합값 증가에 따른 표준화 잔차의 반영되지 못한 분산증가는 산포모형 (2.2)를 통하여 타당하게 모형화 되었음을 의미한다. 표준화된 잔차의 정규분포분위수대조도 그림 2.1 (c)와 히스토그램 그림 2.1 (d)를 그림 1.2의 (c), (d)와 각각 비교하면, 평균과 분산의 동시모형에서의 표준화된 잔차가 더 직선에 가깝고, 대칭적이라는 것을 알 수 있다. 이는 모형 (2.1)과 (2.2)가 모형 (1.1)보다 더 타당한 분석모형이라는 것을 의미한다. 그림 2.2는 산포모형 (2.2)에 대한 잔차그림들로 등분산성을 만족하므로 가정한 산포모형 (2.2)가 적절함을 알 수 있다. 이와 같이 잔차그림들을 이용하면 모형의 타당성을 시각적으로 세심하게 검토할 수 있으므로, 검정통계량으로 모형의 적합도를 검정하기에 앞서 잔차그림을 이용한 모형의 타당성 검토가 먼저 수행되어야 한다고 생각한다.

3. 산포모형 적합도검정을 위한 통계량

Jung과 Lee(1997)는 평균과 분산함수의 타당성을 검정하기 위하여 잔차의 부분합과 잔차제곱과 분산함수와의 차이의 부분합에 근거한 다음의 검정통계량 G를 제안하였다.

$$G = \max\{G_1, G_2\},$$

여기서

$$G_1 = \sup_t |W_1(t)|, \quad G_2 = \sup_t |W_2(t)|,$$

$$W_1(t) = n^{-1/2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i) I(x_i \le t),$$

$$W_2(t) = n^{-1/2} \sum_{i=1}^{n} \{(y_i - \hat{\mu}_i)^2 - \widehat{Var}(y_i)\} I(x_i \le t),$$

이고, $\hat{\mu}_i$과 $\widehat{Var}(y_i)$는 평균과 분산모형에 대한 주어진 가정하에서 추정된 평균과 분산이고, $I(\cdot)$는 표시함수(indicator function), 사건 $(x \le t)$는 t 보다 작거나 같은 x로 구성된다. 검정통계량 G는 평균모형의 적합도를 판정하는 통계량 G_1과 분산모형의 적합도를 판정하는 통계량 G_2중 큰 값으로 모형의 적합도를 검정하기 때문에 분산모형의 타당성을 세심하게 검토하는데는 적절치 못할 수 있다. 모형 (1.1)과 동일한 평균모형을 가정하고 분산함수가 다양한 멱평균모형(power-of-the mean model)일 때 수행된 Jung과 Lee(1997)의 모의실험에서 적합도검정 통계량 G의 경험적 수준(empirical level)은 미리 정해진 명목수준(nominal level)보다 작아 G를 이용한 분산모형의 적합도검정은 보수적임을 알 수 있다. 즉, 평균과 분산 모형을 동시에 검정하므로 평균모형이 타당할 경우 분산모형이 틀리더라도 검정통계량 G의 값이 작아 분산모형이 타당하다고 결론내릴 수 있다.

Lee와 Nelder(1998)가 산포모수의 추론에 사용한 Cox 와 Reid (1987)의 수정단면 확장된 준가능도는 혼합선형모형(mixed linear model)에서의 Patterson과 Thompson(1971)의

표 3.1: 산포모형 (2.2)와 (3.1)에서의 제한가능도 값 및 자유도

산포모형	제한가능도 값	자유도
$\log(\phi_i) = \gamma_0 + \gamma_1 \log(x_i)$	1255.0	104
$\log(\phi_i) = \gamma$	1259.9	105

제한가능도를 비정규분포인 경우로 확장한 것이다. 이 제한가능도는 평균모수의 최대가능도 추정량이 주어졌을 때 계산한 조건부가능도(conditional likelihood)이며, 산포모수만으로 구성된 단면가능도함수(profile likelihood function)이므로 산포모형의 적합도검정은 Lee와 Nelder(1998)의 수정단면 확장된 준가능도 즉, 제한가능도를 사용하여 수행하는 것이 바람직하다. 서로 다른 산포모형에 대한 적합도검정은 자유도와 제한가능도 값의 크기를 각각 비교함으로써 수행되며, 제한가능도 값이 유의하게 작은 산포모형이 타당한 것으로 선택된다.

평균과 분산의 동시모형 (2.1)과 (2.2)에서 산포모형 (2.2) 대신에 다음 모형 (3.1)를 사용하여 분석을 실시하여 보았다.

$$Var(y_i) = \phi_i \mu_i^2, \quad \log(\phi_i) = \gamma. \tag{3.1}$$

모형 (3.1)은 Carroll과 Ruppert(1988)의 분산모형과 동일하다. 표 3.1는 두 산포모형 (2.2)와 (3.1)로 각각 분석했을 때의 제한가능도 값과 자유도를 정리한 것이다. 표 3.1로부터 산포모형 (2.2)는 모형 (3.1) 보다 자유도가 1 감소하지만 제한가능도 값이 4.9 감소하여 유의확률이 0.028이다. 즉, 모형 (2.2)가 모형 (3.1)보다 더 타당한 산포모형임을 알 수 있다. 이와 같이 산포모형의 적합도검정에 제한가능도를 사용하면 평균모수를 제거한 후 산포모수에 대한 적합도검정을 수행하므로 검정과 결과의 해석이 용이하다. 이와 더불어 Lee와 Nelder(1998)의 잔차그림들을 이용하면 모형의 타당성을 시각적으로 확인하면서 검정할 수 있어 산포모형에 대한 보다 신뢰성있는 모형검토를 수행할 수 있겠다.

본 논문의 모형 분석에 사용된 소프트웨어는 전자우편 youngjo@plaza.snu.ac.kr을 통하여 이용가능하다.

참고문헌

[1] Carroll, R.J. and Ruppert, D. (1988). *Transformation and weighting in regression*, Chapman & Hall, New York.

[2] Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate inference, *Journal of the Royal Statistical Society*, B, vol. 49, 1-39.

[3] Jung, S. and Lee, K. (1997). Goodness-of-fit test for mean and variance function, *Journal of the Korean Statistical Society*, vol. 26, 199-210.

[4] Lee, Y. and Nelder, J.A. (1998). Generalized linear models for the analysis of quality-improvement experiments, *The Canadian Journal of Statistics*, vol. 26, 95-105.

[5] McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*, Chapman and Hall, London.

[6] Nelder, J.A. and Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments, *Applied Stochastic Models and Data Analysis*, vol. 7, 107-120.

[7] Nelder, J.A. and Lee, Y. (1997). Letters to the editor, *Technometrics*, vol. 40, 168-175.

[8] Patterson, H.D. and Thompson, D. (1971). Recovery of interblock information when the block size are unequal, *Biometrika*, vol. 58, 545-554.

[1999년 8월 접수, 2000년 5월 채택]

Regression Diagnostics on Joint Modelling of Mean and Dispersion *

Weechang Kang[1] Youngjo Lee[2] Moon Sup Song[3]

ABSTRACT

Carroll and Ruppert(1988) analyzed the esterase assay data with regression model based on quasi-likelihood. Jung and Lee(1997) introduced a goodness-of-fit test for testing the adequacy of the quasi-likelihood and claimed that there is no gross inadequacy with the model because their test was not rejected. However, Lee and Nelder(1998)'s residual plots revealed that the model did not sufficiently reflect the increase of the variance with that of the mean. In this paper, we re-analyze the esterase assay data with the joint modelling of mean and dispersion in Lee and Nelder(1998) and evaluate the validity of the fitted model by applying the residual plots. And it is illustrated that Lee and Nelder(1998)'s restricted likelihood is more efficient in goodness-of-fit test for the dispersion model.

Keywords: Adjusted profile extended quasi-likelihood; Extended quasi-likelihood; Joint modelling of mean and dispersion.

* The present study was supported by the Basic Science Research Institute Program, Ministry of Education, 1998, Project No. 1998-015-D00046.

1) Office for Biostatistics Researches, ASAN Medical Center. E-mail: weechang@www.amc.seoul.kr

2) Associate Professor, Department of Statistics, Seoul National University. E-mail: youngjo@plaza.snu.ac.kr

3) Professor, Department of Statistics, Seoul National University. E-mail: songms@plaza.snu.ac.kr