Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 1 pp. 129~139

A Score Test for Detection of Outliers in Generalized Linear Models [1)]

Myung-Wook Kahng[2)] · Min Kyung Kim[3)]

Abstract

We consider the problem of testing for outliers in generalized linear model. We proceed by first specifying a mean shift outlier model, assuming the suspect set of ourliers is known. Given this model, we discuss standard approaches to obtaining score test for outliers as an alternative to the likelihood ratio test.

Keywords : Generalized linear model, Likelihood ratio test, Outlier, Score test

1. 서론

정규이론에 의한 선형모형은 설명변수의 선형결합(linear predictor)이 직접 반응변수를 설명한다. 이러한 선형모형은 분산분석, 선형회귀분석 등에서 매우 다양한 모형의 틀을 제공하는 것이 사실이지만 모든 상황에서 충분한 것은 아니다. 반응변수가 이항반응자료처럼 이산형 자료인 경우에는 원칙적으로 일반 선형모형은 적절하지 못한 경우가 많고 정규이론에 의한 고전적 선형모형을 확장한 일반화선형모형(generalized linear models)으로 해결이 가능하다.

선형모형의 분석에서는 이상점 검정(outlier test)에 대한 많은 연구가 발표되었고 이를 그래프를 통하여 분석하는 여러 가지 방법도 제시되었으나 일반화선형모형에서는 잔차(residual)와 레버리지(leverage)를 이용하여 이상점을 식별하는 방법 정도만이 알려져 있으며 이러한 방법들은 어느 정도 주관적인 판단을 바탕으로 하기 때문에 객

---

1) This Research was supported by the Sookmyung Women's University Research Grants 2002.

2) First Author : Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea
E-mail : mwkahng@sookmyung.ac.kr

3) Instructor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea

관적으로 이상점을 식별하는데는 어려움이 있다. 따라서 일반화선형모형에서 좀 더 객관적으로 이상점을 검정하는 방법에 대한 연구가 필요하며, 그 한 방법으로 우도비 검정(likelihood ratio test)을 생각해 볼 수 있다.

평균이동이상점모형(mean shift outlier model)을 이용한 이상점 검정에서의 우도비 검정은 대립가설 하에서 최대우도추정량의 계산을 필요로 하는데 선형모형에서와 달리 일반화선형모형에서는 많은 계산이 요구되며 때로는 너무 복잡하여 그것이 회귀진단에서 사용하기에는 적절하지 않은 경우가 발생한다. 이에 대한 해결방법으로 오직 귀무가설 하에서의 최대우도추정량의 계산만이 요구되는 스코어 검정(score test)이 사용될 수 있다. 본 연구에서는 일반화선형모형에서 스코어 검정을 이용한 이상점 검정에 대하여 알아본다.

2. 일반화선형모형에서의 이상점

2.1 일반화선형모형

Nelder and Wedderburn(1972)에 의해 체계화된 일반화선형모형은 지수족(exponential family)분포와 연결함수(link function)를 이용하여 정규이론에 의한 선형모형을 다음과 같은 두 가지 과정으로 일반화한 것이다. 첫째, 오차의 분포는 정규분포를 포함하는 지수족의 여러 가지 분포를 사용한다. 둘째, 반응변수의 기대값과 설명변수의 선형결합을 연결시키는 연결함수를 설정한다. 고전적 선형모형은 반응변수가 서로 독립적이며 정규분포를 따르고 연결함수가 항등함수(identity function)인 일반화선형모형의 특수한 형태라고 할 수 있다.

먼저 반응변수가 다양한 형태의 확률분포를 가질 수 있도록 하기 위해서는 공통된 특성을 갖는 확률분포의 집합을 정의할 필요가 있는데 확률변수 Y의 확률밀도함수가

$$f_Y(y\,;\,\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\} \tag{2.1}$$

의 형태로 표현되고 ϕ가 알려져 있다면 Y의 분포는 지수족에 속한다고 한다. 여기서 $a(\cdot)$, $b(\cdot)$, $c(\cdot)$는 특정 함수이다. 지수족에 대한 좀 더 자세한 내용은 Barndorff and Nilsen(1978)에 설명되어 있다. 실제로 우리가 사용하고 있는 많은 분포들이 지수족에 속하는데 분산이 알려져 있는 경우의 정규분포, 이항분포, 포아송분포, 감마분포 등이 여기에 해당된다.

식(2.1)로 부터 Y에 대한 로그우도함수(log-likelihood function)는 다음과 같이 나타내어진다.

$$l(\theta,\phi\,;\,y) = \log f_Y(y\,;\,\theta,\phi) = (y\theta - b(\theta))/a(\phi) + c(y,\phi) \tag{2.2}$$

Y의 기대값과 분산은 다음의 두 관계식 $E(\partial l/\partial\theta) = 0$과

$E(\partial^2 l/\partial\theta^2) + E(\partial l/\partial\theta)^2 = 0$ 으로부터
$E(Y) = \mu = b'(\theta)$, $Var(Y) = b''(\theta)a(\phi)$가 된다.

2.2 이상점 검정

선형모형에서 중요한 가정의 하나는 자료에 포함된 모든 관측값에 대해 사용된 모형이 적절하다는 것이다. 그러나 실제 문제에서는 한 두개 또는 그 이상의 관측값이 대부분의 관측값이 적합된 모형에 일치하지 않는 경우를 많이 볼 수 있다. 이와 같이 자료의 나머지 관측값들과는 다르게 주어진 모형을 따르지 않는 관측값을 이상점(outlier)이라고 한다. 이상점들은 분석시 포함여부에 따라 모형의 근본 모양이 바뀔 수도 있고, 어떠한 경우에는 대다수의 관측값들이 가지고 있지 않는 중요한 특징을 갖는 이상점을 찾아내는 것이 오히려 분석의 목표가 될 수도 있다.

관찰점들의 개수가 적은 경우에는 이상점을 찾기 위해서 육안으로 관측값들을 살펴보거나 산점도 등을 사용할 수 있다. 그러나 관찰점들의 개수가 많거나 자료에 많은 설명변수들이 있다면 이러한 방법들은 별로 유용하지 못하다. 더구나 이러한 방법들은 주관적인 판단을 바탕으로 하기 때문에 객관적으로 이상점을 식별하는데 어려움이 있다(Beckman and Cook, 1983).

고전적 선형모형에서와 마찬가지로 일반화선형모형에서 이상점을 식별할 수 있는 두 가지 대표적인 통계량으로 레버리지와 잔차를 들 수 있다(McCullagh and Nelder, 1989). 레버리지는 관측값이 설명변수 공간의 중심에서 얼마나 멀리 떨어져 있는지의 여부를 평가하는 통계량이고 잔차는 관측값과 예측값 사이에 얼마나 큰 차이가 있는지를 평가하는 통계량이다.

선형모형에서의 레버리지는 햇행렬(hat matrix) $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$의 대각원소로 정의하였다. 일반화선형모형에서 햇행렬 $\boldsymbol{H}^*$는 $\boldsymbol{H}^* = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$이다. 여기서, $\boldsymbol{W} = diag\{w_i\}$이고 w_i은 가중치로서 다음절의 식 (3.2)에 설명된다. 따라서 일반화선형모형에서 i번째 자료의 레버리지는 $\boldsymbol{H}^*$행렬의 대각원소인 h_{ii}^*로 정의된다. p개의 설명변수를 포함하는 자료에서 일반적으로 i번째 레버리지값이 레버리지의 산술평균인 $\overline{h}^* = (p+1)/n$의 두배 이상이면 i번째 자료는 이상점으로 판정한다. 고전적 선형모형에서는 레버리지값이 클수록 자료의 중심에서 멀어지고 레버리지값이 작을수록 자료의 중심에 가깝게 위치한다는 의미로 해석되었다. 그러나 일반화선형모형에서는 항상 이런 해석이 적용될 수는 없다.

피어슨잔차(Pearson residual)는 $r_i = (y_i - \mu_i)/\sqrt{V(\mu_i)}$로 정의된다. 피어슨잔차는 포아송분포에서 모형의 적합도 평가에 사용되는 피어슨의 카이제곱통계량과 $X^2 = \sum_{i=1}^n r_i^2$인 관계를 가지기 때문에 r_i를 피어슨잔차라고 부르게 되었다. 표준화 피어슨잔차(studentized Pearson residual) r_i^s는 다음과 같이 정의된다.

$$r_i^s = \frac{r_i}{\sqrt{1-h_{ii}^*}} \tag{2.3}$$

McCullagh and Nelder(1989)에 의하면 이상점의 식별을 위해서 일반적으로 피어슨 잔차 r_i보다 레버리지에 의해 잔차를 수정한 표준화 피어슨잔차 r_i^s가 더 선호되는데 레버리지 값이 큰 관측값에 대해서 표준화 피어슨잔차를 사용하면 피어슨잔차보다 큰 값을 가지게 되기 때문에 이상점을 더 쉽게 탐지 할 수 있다. 일반적인 기준은 표준화 피어슨잔차의 절대값이 2보다 큰 경우에 이상점일 가능성이 있다고 판정한다.

완전모형(full model) 하에서 $\boldsymbol{\beta}$에 대한 최우추정량을 $\widehat{\boldsymbol{\beta}}$라고 하고, 축소모형(reduced model) 하에서 $\boldsymbol{\beta}$에 대한 최우추정량을 $\boldsymbol{b}$라고 하자. Nelder and Wedderburn(1972)는 모형의 적합도를 판단하는 값으로 데비언스 $D = 2[L(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}) - L(\boldsymbol{b}; \boldsymbol{y})]$를 제안하였다. 여기서 $L(\boldsymbol{\beta}; \boldsymbol{y})$는 로그우도 함수이다. 데비언스 D는 근사적으로 자유도가 $n-p$인 카이제곱 분포를 따른다고 알려져 있다. 데비언스 D는 n개의 항의 합으로 표현된다. 즉 $D = \sum_{i=1}^n d_i$이다. 데비언스잔차 r_k^d는 다음과 같이 정의된다.

$$r_i^d = sign(y_i - \mu_i)\sqrt{d_i} \tag{2.4}$$

데비언스잔차는 적합된 모형에서의 우도의 i번째 구성원소와 완전모형에서의 우도의 대응원소와의 불일치 정도를 측정하게 되므로 큰 값을 가질수록 이상점일 가능성이 커지게 된다.

2.3 이상점 모형

일반화선형모형은 다음과 같이 나타낼 수 있고

$$\eta_i = g(\mu_i) = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \cdots, n$$

행렬을 이용하면 다음과 같이 표현된다.

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X\beta}$$

여기서 $\boldsymbol{\eta}$는 $n \times 1$ 행렬이고 $g(\cdot)$은 연결함수, $\boldsymbol{X}$는 $n \times p$, $\boldsymbol{\beta}$는 $p \times 1$ 행렬이다. 그리고 선형결합 $\boldsymbol{\eta}$는 연결함수 $g(\cdot)$에 의해 $\boldsymbol{\mu} = E(\boldsymbol{Y})$와 연결된다.

m개의 관측값이 이상점이라고 가정하면 벡터 $\mathbf{I} = (i_1, i_2, \cdots, i_m)$를 사용하여 다음과 같은 모형을 가정할 수 있다. 이상점에 대한 모형은 여러 가지 방법으로 만들 수 있지만 여기서는 평균이동이상점모형을 생각해 보겠다(Cook and Weisberg, 1982).

$$\eta_i = \begin{cases} \sum \beta_j x_{ij}, & i \notin \mathrm{I} \\ \sum \beta_j x_{ij} + \delta_i, & i \in \mathrm{I} \end{cases}$$

행렬을 사용하여 위의 모형을 나타내면

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{D\delta}$$

이고, 여기서 $\boldsymbol{\delta} = (\delta_{i_1}, \delta_{i_2}, \cdots, \delta_{i_m})^T$, $\boldsymbol{D} = (\boldsymbol{d}_{i_1}, \boldsymbol{d}_{i_2}, \cdots, \boldsymbol{d}_{i_m})$이고 $\boldsymbol{d}_{i_j}$는 i_j번째 원소만 1의 값을 가지고 나머지 원소는 0인 벡터이다. 그러므로 이 경우 $\boldsymbol{\delta} = \mathbf{0}$에 대한 검정은 바로 m개의 관측값이 이상점인지를 판단하는 검정이 된다.

3. 스코어 검정

Rao(1947)가 처음 제안하고 Silvey(1959)에 의해 체계화된 스코어 검정은 우도비 검정에서의 계산상의 복잡함을 해결하는 대안으로 널리 사용되어져 왔다. 또한 스코어 검정은 라그랑지승수검정(Lagrange multiplier test)이라고 부르기도 한다. 가설 H_0: $\boldsymbol{\phi} = \boldsymbol{\phi}_0$에 대한 스코어 통계량의 일반적인 형태는 다음과 같다.

$$S = \boldsymbol{U}(\boldsymbol{\phi}_0)^T \boldsymbol{I}(\boldsymbol{\phi}_0)^{-1} \boldsymbol{U}(\boldsymbol{\phi}_0) \tag{3.1}$$

여기서, 효율스코어(efficient score)벡터 $\boldsymbol{U}$와 정보행렬(information matrix) $\boldsymbol{I}$는 다음과 같다.

$$\boldsymbol{U}(\boldsymbol{\phi}) = \frac{\partial L(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}, \quad \boldsymbol{I}(\boldsymbol{\phi}) = -E\left(\frac{\partial^2 L(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\, \partial \boldsymbol{\phi}^T}\right).$$

우리의 관심이 $\boldsymbol{\delta}$에 있다고 하고 $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\delta})$로 놓으면 효율스코어벡터는

$$\boldsymbol{U}(\boldsymbol{\phi}) = \boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \left[\left(\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\beta}}\right)^T \left(\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)^T\right]^T$$

이고, 정보행렬은

$$I(\phi) = I(\beta, \delta) = \begin{bmatrix} I_{11}(\beta, \delta) & I_{12}(\beta, \delta) \\ I_{21}(\beta, \delta) & I_{22}(\beta, \delta) \end{bmatrix} = \begin{bmatrix} -E\frac{\partial^2 L(\beta, \delta)}{\partial \beta \partial \beta^T} & -E\frac{\partial^2 L(\beta, \delta)}{\partial \beta \partial \delta^T} \\ -E\frac{\partial^2 L(\beta, \delta)}{\partial \delta \partial \beta^T} & -E\frac{\partial^2 L(\beta, \delta)}{\partial \delta \partial \delta^T} \end{bmatrix}$$

로 나타내어진다.

식 (2.2)에 의하면 i번째 관측값에 대한 로그우도는 다음과 같고,

$$l(\theta_i; y_i) = l_i = (y_i \theta_i - b(\theta_i)) / a(\phi) + c(y_i, \phi)$$

n개의 관측값에 대한 로그우도함수는 다음과 같이 나타낼 수 있다.

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \{ (y_i \theta_i - b(\theta_i)) / a(\phi) + c(y_i, \phi) \}$$

그리고 $\partial l_i / \partial \beta_j$ 는 연쇄법칙(chain rule)에 의해 다음과 같이 나타내어진다.

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

여기서 $b'(\theta_i) = \mu_i$이고, $Var(y_i) = \sigma_i^2$라 하면,

$$\frac{d\mu_i}{d\theta_i} = b''(\theta_i) = \frac{Var(y_i)}{a(\phi)} = \frac{\sigma_i^2}{a(\phi)}$$

가 된다. 또, $\eta_i = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}$로 부터 $d\eta_i / d\mu_i = g'(\mu_i)$, $\partial \eta_i / \partial \beta_j = x_{ij}$를 얻을 수 있다. 그러면 $\partial l_i / \partial \beta_j$ 는 다음과 같이 계산된다.

$$\begin{aligned} \frac{\partial l_i}{\partial \beta_j} &= \frac{\{y_i - b'(\theta_i)\}}{a(\phi)} \frac{a(\phi)}{Var(y_i)} \frac{d\mu_i}{d\eta_i} x_{ij} \\ &= (y_i - \mu_i)\, w_i \frac{d\eta_i}{d\mu_i} x_{ij} \\ &= (y_i - \mu_i)\, w_i\, g'(\mu_i)\, x_{ij} \end{aligned}$$

여기서 가중치 w_i는

$$w_i^{-1} = \left(\frac{d\eta_i}{d\mu_i}\right)^2 Var(y_i) = \left(\frac{d\eta_i}{d\mu_i}\right)^2 \sigma_i^2 \tag{3.2}$$

로 정의한다. 따라서 n개의 관측값의 경우로 확장하면 $\partial L / \partial \beta_j$ 는 다음과 같다.

$$\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i)\ w_i\, g'(\mu_i)\ x_{ij}$$

$\partial l_i / \partial \delta_k$ 도 같은 방법으로 다음과 같이 구할 수 있다.

$$\begin{aligned}\frac{\partial l_i}{\partial \delta_k} &= \frac{\partial l_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \delta_k} \\ &= (y_i - \mu_i)\ w_i\, g'(\mu_i)\ d_{ik}\end{aligned}$$

여기서 d_{ik}는 벡터 $\boldsymbol{d}_k$의 i번째 원소를 나타내며 벡터 $\boldsymbol{d}_k$는 k번째 원소만 1이고 나머지 원소는 모두 0인 벡터이다. 그러므로 n개의 관측값에 대해 $\partial L / \partial \delta_k$는 다음과 같다.

$$\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \delta_k} = \sum_{i=1}^{n} (y_i - \mu_i)\ w_i\, g'(\mu_i)\ d_{ik} = (y_k - \mu_k)\ w_k\, g'(\mu_k)$$

이제 로그우도를 $\boldsymbol{\beta}$에 대해서 이차 미분하면

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \beta_j \partial \beta_s} &= \frac{\partial}{\partial \beta_s}\left[\sum_{i=1}^{n} (y_i - \mu_i)\ w_i \frac{d\eta_i}{d\mu_i} x_{ij}\right] \\ &= \sum_{i=1}^{n}\left[(y_i - \mu_i)\frac{\partial}{\partial \beta_s}\left\{w_i \frac{d\eta_i}{d\mu_i} x_{ij}\right\} + w_i \frac{d\eta_i}{d\mu_i} x_{ij} \frac{\partial}{\partial \beta_s}(y_i - \mu_i)\right]\end{aligned}$$

이고, 여기에 기대값을 취하면 $E(y_i) = \mu_i$이므로 첫 번째 항은 소거되고 두 번째 항만 남아 다음과 같이 계산된다.

$$-E\left[\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \beta_j \beta_s}\right] = \sum_{i=1}^{n} w_i \frac{d\eta_i}{d\mu_i} x_{ij} \frac{\partial \mu_i}{\partial \beta_s} = \sum_{i=1}^{n} w_i x_{ij} x_{is}$$

마찬가지로 $\boldsymbol{\delta}$에 대해 이차 미분하면 다음과 같고

$$\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \delta_k \partial \delta_r} = \frac{\partial}{\partial \delta_r}\left[(y_k - \mu_k)\, w_k \frac{d\eta_k}{d\mu_k}\right]$$
$$= \left[(y_k - \mu_k)\frac{\partial}{\partial \delta_r}\left\{w_k \frac{d\eta_k}{d\mu_k}\right\} + w_k \frac{d\eta_k}{d\mu_k}\frac{\partial}{\partial \delta_r}(y_k - \mu_k)\right]$$

기대값을 취하면

$$-E\left[\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \delta_k \delta_r}\right] = w_k \frac{d\eta_k}{d\mu_k}\frac{\partial \mu_k}{\partial \delta_r} = \begin{cases} 0 \ , & r \neq k \\ w_k \ , & r = k \end{cases}$$

이다. 또, $\boldsymbol{\beta}$와 $\boldsymbol{\delta}$에 대해 미분을 하고 기대값을 취하면 다음과 같다.

$$-E\left[\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \beta_j \partial \delta_k}\right] = \sum_{i=1}^{n} w_i \frac{d\eta_i}{d\mu_i} x_{ij} \frac{\partial \mu_i}{\partial \delta_k} = \sum_{i=1}^{n} w_i x_{ij} d_{ik} = w_k x_{kj}$$

귀무가설 $\boldsymbol{\delta} = 0$ 하에서 행렬을 이용하면 효율스코어 벡터 $\boldsymbol{U}(\boldsymbol{\phi}_0)$ 는

$$\boldsymbol{U}(\boldsymbol{\phi}_0) = \boldsymbol{U}(\boldsymbol{\beta}, 0) = \begin{bmatrix} 0 \\ \boldsymbol{a} \end{bmatrix}$$

이다. 여기서 $\boldsymbol{a} = \{(y_{i_j} - \mu_{i_j})\, w_{i_j} g'(\mu_{i_j})\}$인 $m \times 1$ 벡터이고, 정보행렬 $\boldsymbol{I}(\boldsymbol{\phi}_0)$ 는

$$\boldsymbol{I}(\boldsymbol{\phi}_0) = \begin{bmatrix} \boldsymbol{I}_{11}(\boldsymbol{\beta}, 0) & \boldsymbol{I}_{12}(\boldsymbol{\beta}, 0) \\ \boldsymbol{I}_{21}(\boldsymbol{\beta}, 0) & \boldsymbol{I}_{22}(\boldsymbol{\beta}, 0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{D} \\ \boldsymbol{D}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D} \end{bmatrix}$$

로 나타낼 수 있다.

계산상의 편의를 위해 정보행렬의 역행렬 $\boldsymbol{I}(\boldsymbol{\phi}_0)^{-1}$을 다음과 같이 두면,

$$\boldsymbol{I}(\boldsymbol{\phi}_0)^{-1} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}$$

이고, 귀무가설 하에서 스코어 통계량 (3.1)은

$$S = \boldsymbol{U}(\boldsymbol{\phi}_0)^T \boldsymbol{I}(\boldsymbol{\phi}_0)^{-1} \boldsymbol{U}(\boldsymbol{\phi}_0) = \boldsymbol{a}^T \cdot \boldsymbol{B}_{22} \cdot \boldsymbol{a}$$

으로 나타내어진다. $\boldsymbol{X}^* = \boldsymbol{W}^{1/2}\boldsymbol{X}$, $\boldsymbol{D}^* = \boldsymbol{W}^{1/2}\boldsymbol{D}$, $\boldsymbol{H}^* = \boldsymbol{X}^*(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}$ 라고 하면,

$$\begin{aligned}\boldsymbol{B}_{22} &= [\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D} - \boldsymbol{D}^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{D}]^{-1} \\ &= [\boldsymbol{D}^{*T}(\boldsymbol{I} - \boldsymbol{H}^*)\boldsymbol{D}^*]^{-1}\end{aligned}$$

이고, 따라서 스코어 통계량은 다음과 같다.

$$S = \boldsymbol{a}^T \cdot \boldsymbol{D}^{*T}(\boldsymbol{I} - \boldsymbol{H}^*)\boldsymbol{D}^* \cdot \boldsymbol{a}$$

만약 k번째 관측값 하나만 이상점이라 하면, $\boldsymbol{D}^{*T}(\boldsymbol{I} - \boldsymbol{H}^*)\boldsymbol{D}^* = w_k(1 - h^*_{kk})$이므로 k번째 관측값에 대한 스코어 통계량은 다음과 같다.

$$S_k = \frac{\{(y_k - \hat{y}_k) \cdot w_k \cdot 1\}^2}{w_k(1 - h^*_{kk})} \tag{3.3}$$

여기서 h^*_{kk} 는 행렬 $\boldsymbol{H}^*$의 k번째 대각원소이다. 또한 정규분포의 가정에서 k번째 관측값이 이상점이라면 스코어 통계량은 다음과 같다.

$$S_k = \frac{{e_k}^2}{\sigma^2(1 - h_{kk})}$$

여기서 h_{kk}는 행렬 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$의 k번째 대각원소이다. 위의 결과에서 알 수 있듯이 반응변수가 정규분포를 따를 경우 스코어 통계량은 일반 선형모형에서의 스튜던트화잔차와 같은 형태가 됨을 알 수 있다.

만약 n개의 관측값 중 m개의 이상점이 있다고 했을 때, 우리는 m개의 원소를 가지고 있는 ${}_n\mathrm{C}_m$개의 부분집합 $\mathbf{I} = (i_1, i_2, \cdots, i_m)$을 생각해야된다. 그 중에서 가장 큰 검정통계량을 가진 부분집합이 이상점일 가능성이 가장 높은 부분집합이 될 것이다. 만약 우리가 우도비검정을 실시하고자 한다면 ${}_n\mathrm{C}_m$개의 모든 부분집합에 대해서 우도비검정통계량을 구해야하고 그것은 같은 회수의 일반화선형모형의 적합을 필요로 한다. 비록 n이 그리 크기 않더라도 m이 2나 3정도만 되어도 우리가 적합시키고 계산해야될 부분은 매우 많아져서 이것을 회귀진단의 도구로 사용하는데는 큰 무리가 따른다. 그러나 스코어검정을 이상점 검토의 기준으로 한다면 n이나 m의 크기에 관계없이 단 한번의 적합만으로 모든 통계량의 계산이 가능하므로 이 방법이 회귀진단의 도구로서 매우 적절하다고 할 수 있다.

4. 예 제

1996년 챌린저호의 사고 이전에 일어난 23대의 우주왕복선 사고 당시의 기온과 주요 O-링이 기온에 의해 손상되었는지 여부를 나타낸 자료가 <표 1>에 제시되어 있다. 손상된 경우는 1, 그렇지 않은 경우는 0이다. 반응변수가 0 또는 1의 값만 가지는 이산형 변수이므로 베르누이 시행을 따르게 된다. 따라서 연결함수로 로짓(logit) 함수를 가지는 로지스틱모형(logistic model)을 사용하여 자료를 적합시켰다.

각각의 관측값들에 대한 레버리지 h^*_{kk}, 식 (2.3)의 표준화 피어슨잔차 r^s_k, 식 (2.4)의 데비언스잔차 r^d_k, 식 (3.3)의 스코어 통계량 S_k을 SAS를 이용하여 구하고 <표 1>에 제시하였다. 레버리지의 기준으로 보면 $\overline{h}^* = 0.087$의 2배인 0.174보다 보다 매우 큰 값을 가지는 9번째와 23번째 값이 영향력이 큰 관측값(influential observation)일 가능성이 높다고 생각할 수 있다. 또한 표준화 피어슨잔차의 경우 절대값이 2보다 큰 21번째 값을 이상점으로 판단할 수 있다. 데비언스잔차의 경우도 21번째 값이 이상점으로 의심이 된다. 스코어통계량도 역시 21번째 값이 11.4840으로 귀무가설 $\delta = 0$ 을 기각하므로 이상점이라고 판단할 수 있다. 이항분포에서는 추정된 확률이 극단적인 경우에 항상 큰 레버리지값이 이상점을 의미하는 것은 아니므로 위의 결과들을 종합해 볼 때 21번째 값이 이상점으로 판단된다.

<표 1> 각 관측값들에 대한 이상점 검정을 위한 통계량

Ft	Temp	TD	h^*_{kk}	r^s_k	r^d_k	S_k
1	66	0	0.08639	-0.90961	-1.50065	0.8274
2	70	1	0.06275	1.89013	2.42472	3.5726
3	69	0	0.06334	-0.63417	-1.13081	0.4022
4	68	0	0.06692	-0.71359	-1.24699	0.5092
5	67	0	0.07444	-0.80467	-1.37049	0.6475
6	72	0	0.06599	-0.44831	-0.82954	0.2010
7	73	0	0.06786	-0.39958	-0.74496	0.1597
8	70	0	0.06275	-0.56448	-1.02240	0.3186
9	57	1	0.20777	0.45459	0.77877	0.2067
10	63	1	0.14294	0.87704	1.42319	0.7692
11	70	1	0.06275	1.89013	2.42472	3.5726
12	78	0	0.06311	-0.22306	-0.42691	0.0498
13	67	0	0.07444	-0.80467	-1.37049	0.6475
14	53	1	0.16748	0.27874	0.50071	0.0777
15	67	0	0.07444	-0.80467	-1.37049	0.6475
16	75	0	0.06915	-0.31701	-0.59808	0.1005
17	70	0	0.06275	-0.56448	-1.02240	0.3186
18	81	0	0.05087	-0.15645	-0.30309	0.0245
19	76	0	0.06814	-0.28211	-0.53495	0.0796
20	79	0	0.05944	-0.19823	-0.38101	0.0393
21	75	1	0.06915	3.38880	3.13606	11.4840
22	76	0	0.06814	-0.28211	-0.53495	0.0796
23	58	1	0.20898	0.51094	0.86654	0.2611

Ft, flight number; Temp, temperature (^oF); TD, thermal distress(1, yes; 0, no).
출처: Dalal(1989)

참고문헌

1. Bandorff-Neilsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, Chichester, Wiley.
2. Beckman, R. J. and Cook, R. D. (1983), Outlier...s, *Technometrics*, Vol. 25, 119-149.
3. Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, London, Chapman Hall.
4. Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989), Risk analysis of the space shuttle: Pre-*Challenger* prediction of failure, *Journal of the American Statistical Association*, Vol. 84, 945-957.
5. McCullage, P. and Nelder, J. A. (1989), *Generalized Linear Models, 2nd ed.*, London, Chapman Hall.
6. Nelder, J. A. and Wedderburn, R. W. M. (1972), Generalized linear models, *Journal of Royal Statistical Society*, A, Vol. 135, 370-384.
7. Rao, C. R. (1947), Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proc. Camb. Phil. Soc.*, Vol. 44, 50-57.
8. Silvey, S. D. (1959), The Lagrangian multiplier test, *Annals of Mathematical statistics*, Vol. 30, 389-407.

[2003년 1월 접수, 2004년 2월 채택]