

Testing Homogeneity of Errors in Unbalanced Random Effects Linear Model

Chul H. Ahn¹⁾

Abstract

A test based on score statistic is derived for detecting homoscedasticity of errors in unbalanced random effects linear model. A small simulation study is performed to investigate the finite sample behaviour of the test statistic which is known to have an asymptotic chi-square distribution under the null hypothesis.

Keywords : score test, constant variance, maximum likelihood estimate

1. 서론

오차에 대한 등분산 가정은 통계학의 많은 분야에서 사용되는 중요한 가정 중 하나라고 할 수 있다. 선형회귀모형에서 오차에 대한 등분산 가정은 정규성에 대한 가정보다 그 중요성이 더욱 강조된다고 할 수 있다. 최근에는 고정효과와 임의효과를 모두 갖는 선형모델에서도 이러한 등분산과 관련된 진단문제가 매우 활발히 논의되어 왔었다. 이러한 예는 le Cessie, S 와 van Houwelingen (1995) 그리고, Jacqmin-Gadda 와 Commenges (1995)에서 찾아 볼 수 있다. 이 논문에서는 선형혼합모형에 있어서 오차에 대한 등분산 검정문제를 다룰 것이다. 오차가 등분산을 갖지 못할 때 오차의 분산은 설명변수들 중 하나 또는 그 이상의 설명변수들과 함수관계를 갖거나 또는 시간이나 공간과 같은 변수들과도 함수관계를 가질 수 있다.

이 논문에서는 이러한 등분산과 관련된 문제를 해결하기 위해 스코어검정법을 유도할 것이다. 스코어방법은 추정이 귀무가설 아래에서만 이루어지고 잘 정돈된 방법으로서 많은 분야에서 효과적인 검정법으로 사용되어 왔었다. 선형회귀분야에서는 Cook 과 Weisberg (1983)가 오차분산의 로그를 설명변수들의 선형함수로 표현하고 이를 이용하여 오차분산의 등분산 여부를 검정하는 스코어검정통계량을 개발하였다. Chi 와 Reinsel (1989)은 조건부독립 임의효과모형에 있어서 그룹내 오차들 사이에 자기상관이 존재하는지 여부를 알아내기 위한 스코어검정통계량을 개발하였다. 일반화선형모형에서는 Smyth (1989)가 분산에 대해 로그링크를 갖는 감마일반화선형모형을 가정하고 등분산여부를 찾아내는 스코어검정법을 제시하였다. 최근에는 Ahn (2000)에서 임의효과에 대한 등분산 여부를 검정하는 스코어통계량을 제시하였다.

이 논문에서는 오차의 분산에 대한 확장모형을 수립하고 이를 기반으로 하여 등분산검정을 위

1) Associate Professor, Applied Mathematics, Sejong University, 98 Gunja-Dong, Gwangjin-Gu, Seoul 143-747, Republic of Korea
E-mail : chahn@sejong.ac.kr

한 스코어검정통계량을 유도한다. 또한, 스코어검정통계량의 귀무가설 분포에 대한 χ^2 근사의 타당성을 알아보기 위해 균형자료와 불균형자료에 대해서 시뮬레이션이 시행된다.

2. 모델 확장

본 논문에서 다를 선형혼합모형은 임의효과를 갖는 선형모형의 특수한 형태이다.

$$Y_{ij} = X_{ij}^T \beta + b_i + \varepsilon_{ij} \quad (1)$$

위 모형은 임의선형모델인 $Y_{ij} = \mu + b_i + \varepsilon_{ij}$ 에서 μ 가 조건부기대치 또는 X_{ij} 의 회귀함수로 표현된다는 것이다. 여기서 Y_{ij} 는 i 번째 그룹의 j 번째 관측값을 나타낸다. 총 그룹수는 t 이고 i 번째 그룹의 관측값 갯수는 n_i 이고 총 관측값의 갯수는 N 이다. X_{ij} 는 $p \times 1$ 벡터로서 Y_{ij} 의 설명변수이고 그 구성요소는 이미 아는 값들이다. β 는 $p \times 1$ 벡터이고, 절편을 포함하고 있을 경우는 $(p+1) \times 1$ 벡터로서 미지의 회귀모수이고 고정효과이다. 서로 독립인 b_i 는 임의효과를 나타내며, b_i 는 평균 0, 분산 σ_b^2 을 갖는 정규분포를 따르는 것으로 가정하고 있다. ε_{ij} 는 서로 독립이고, 그 각각은 평균 0, 분산 σ^2 을 갖는 정규분포를 따르는 것으로 가정된다. 또한, 임의효과 b_i 와 오차 ε_{ij} 는 서로 독립으로 가정된다. 선형혼합모형에서 한가지 중요한 가정은 각각의 오차들이 똑같은 정규모집단에서 나왔다고 하는 것이다. 이제부터 이 가정의 유효함을 알아보기 위한 진단방법을 찾아 보기로 하자. 한가지 방법은 모델확장을 통하는 것이다. 즉, 모형 (1)에 모수를 추가하여 확장된 모형을 만들고 이 추가된 모수에 대해 스코어검정을 실시하는 것이다. 모형 (1)의 오차에 대해 아래의 모델확장을 생각해 볼 수 있다.

$$var(\varepsilon_{ij}) = \sigma^2 w(z_i, \lambda) \quad (2)$$

여기서 z_i 는 그룹 i 에 대한 $q \times 1$ 공변수벡터이고, 구성요소는 z_{ik} 로서 아래인자 k 는 1에서 q (q 는 공변수의 갯수)까지의 값을 취한다. λ 는 $q \times 1$ 벡터의 미지 모수이다. 함수 $w(\cdot)$ 는 일종의 가중함수로 생각되어 질 수 있으며, λ 에 대해 두번 미분 가능한 함수로서 모든 z_i 에 대해 $w(\lambda_0^T z_i) = 1$ 를 만족하는 특정한 값인 λ_0 가 존재한다고 가정한다. 따라서 등분산을 위한 검정은 자연히 $\lambda = \lambda_0$ 가 될 것이다. 함수 w 가 취할 수 있는 형태중 가장 유용한 것은 지수함수 일 것이다. 왜냐하면, 지수함수는 계속 미분 가능하고 $\lambda = 0$ 일때 등분산을 회복하게 되기 때문이다. 뿐만이 아니라 지수는 단조함수이므로 분산이 어떤 방향으로 증가하는 형태를 보일 경우, 그 방향을 쉽게 찾을 수 있다는 것이다. 즉, 분산이 증가하는 방향은 $\lambda^T z_i$ 일 것이다. $w(\cdot)$ 의 형태로 본 논문에서 고려하고 있는 함수는 다음과 같다.

$$w(z_i, \lambda) = \exp\left(\sum_{k=1}^q \lambda_k z_{ik}^{\alpha_k}\right) \quad (3)$$

여기서 z_{ik} 는 z_i 의 요소이며, z_i 는 앞에서 정의되었듯이 그룹 i 를 위한 공변수 벡터로서 대개는 설명변수인 X 의 행렬에서 취해진다. 위 식 (3)은 벡터 z_i 가 λ 에 곱해지기 전에 제곱이나 로그 변환을 취할 때 이분산(heteroscedasticity)이 탐지될 수도 있음을 반영하고 있다. (3)에서 $\alpha_k = 0$ 은 로그변환을 가리킨다. 분산이 반응변수의 기대값에 따라 달라질 때에는 w 가 다음과 같은 함수형태를 갖는 것이 유용할 것이다.

$$w(\lambda^t z_i) = w(E(y_i)) \quad (4)$$

식(2)에 있는 확장모델이 암시하고 있는 것은 다른 그룹에 속한 반응변수들이 서로 다른 분산을 갖고 있다는 것을 의미하며, 관측치들이 큰 그룹의 오차분산이 크다는 것을 나타내고 있으며 이러한 분산의 크기는 z_i 를 통하여 모형화할 수 있다는 것이다.

3. 스코어검정통계량

오차의 등분산을 위한 검정은 가중함수(weight function)인 (3)에서 $\lambda = \lambda_0$ 이 될 것이다. 이제, 귀무가설, $H_0: \lambda = \lambda_0$ 와 대립가설, $H_1: \lambda \neq \lambda_0$ 을 검정하기 위한 스코어검정통계량을 유도한다. 먼저, 조금 더 일반적인 경우를 생각해 보기 위해 확률밀도함수 $f(y; \theta)$ 를 갖는 확률벡터 y 를 생각해보자. θ 는 $\theta \in \Theta \subseteq R^r$ 로 표시되는 $r \times 1$ 모수벡터이다. 우리 문제의 경우, θ 는 $\beta, \sigma^2, \sigma_b^2, \lambda$ 를 포함할 것이다. 확률밀도함수 $f(y; \theta)$ 는 Serfling (1980, p. 144)의 정규조건(regularity conditions)을 만족한다고 가정한다. 확률벡터 y 에서 얻어지는 t 개의 독립적인 관측치벡터는 y_1, y_2, \dots, y_t 로 표시하기로 하자. 그러면, 확률벡터 y_1, y_2, \dots, y_t 들의 로그우도함수인 $I(\theta)$ 는 $I(\theta) = I_1(\theta) + I_2(\theta) + \dots + I_t(\theta)$ 로 표시된다. 여기서, $I_i(\theta) = \log f(y_i; \theta)$ 이다. 이제 θ 가 θ_1 과 θ_2 로 분할된다고 하자. 우리 문제의 경우, θ_1 은 $(\beta, \sigma^2, \sigma_b^2)$ 를 포함하고, θ_2 는 λ 를 포함할 것이다. 즉, $\theta = (\theta_1^t, \theta_2^t)^t$. 그리고, θ_2 는 $q \times 1$ 모수벡터라고 하자. 이제, 스코어벡터와 정보행렬의 구성요소는 각각 다음과 같이 쓰여진다.

$$d_j = \sum_i \partial I_i(\theta) / \partial \theta_j, \quad J_{jk} = -E[\sum_i \partial^2 I_i(\theta) / \partial \theta_j \partial \theta_k^t].$$

여기서 \sum 의 인자 i 는 1에서 t 까지의 값을 취하고, 인자 j 와 k 는 1과 2를 취한다. d_j 과 J_{jk} 를 각각 d_j 과 J_{jk} 이 $\theta_1 = \hat{\theta}_1$ ($\theta_2 = 0$ 에서 최우추정량) 일 때 평가된 통계량이라 하자. 이제, 귀무가설, $H_0: \theta_2 = 0$ 와 대립가설, $H_1: \theta_2 \neq 0$ 을 검정하는 스코어검정통계량은 Cox 과

Hinkley (1974, 9장)에 의하면 다음과 같이 쓰여질 수 있다.

$$S = \hat{d}_2^t (\hat{J}_{22} - \hat{J}_{21} \hat{J}_{11}^{-1} \hat{J}_{12})^{-1} \hat{d}_2 \quad (5)$$

스코어검정통계량 S 는 귀무가설이 $H_0: \theta_2 = 0$ 일 때 점근적으로 자유도 q 를 갖는 χ^2 분포를 따른다. 스코어검정은 점근적으로 우도비검정과 같다. 다만, 스코어검정은 귀무가설에서만 추정이 이루어지면 검정통계량이 계산되었으나, 우도비검정은 대립가설에서도 추정이 이루어져야 한다는 점에서 종종 스코어검정이 선호되기도 한다. 이제, 분산에 대한 확장모형 (2)를 생각해 보자. 그룹 i 의 반응벡터 y_i 가 갖는 분산-공분산행렬은 다음과 같이 쓰여질 수 있다.

$$Cov(y_i) = \sigma^2 w(\lambda^t z_i) I_i + \sigma_b^2 \mathbf{1}_i \mathbf{1}_i^t \quad (6)$$

여기서 y_i 는 $n_i \times 1$ 벡터이며, I_i 는 $n_i \times n_i$ 단위행렬이고, $\mathbf{1}_i$ 는 $n_i \times 1$ 의 열행렬로서 모두 1 의 값을 갖는다. 이제 ξ 를 두 분산, σ_b^2 와 σ^2 의 비율로 표시하면 분산-공분산행렬, $Cov(y_i)$ 은 다음과 같이 쓰여진다. 즉, $Cov(y_i) = \sigma^2 Q_i$ 이고, 여기서 $Q_i = w(\lambda^t z_i) I_i + \xi \mathbf{1}_i \mathbf{1}_i^t$. 그리고 우도함수는 다음과 같이 주어진다.

$$l_i(\theta, \lambda) = -\frac{n_i}{2} \log 2\pi - \frac{1}{2} \log |\sigma^2 Q_i| - \frac{1}{2} \varepsilon_i^t (\sigma^2 Q_i)^{-1} \varepsilon_i \quad (7)$$

여기서 $\theta = (\beta^t, \sigma^2, \xi)^t$, 그리고 $\varepsilon_i = y_i - X_i^t \beta$. 이전, 귀무가설, $H_0: \lambda = 0$ 와 대립가설, $H_1: \lambda \neq 0$ 을 검정하기 위한 스코어검정통계량은 다음과 같다.

$$S = \hat{d}_\lambda^t (\hat{J}_{\lambda\lambda} - \hat{J}_{\lambda\theta} \hat{J}_{\theta\theta}^{-1} \hat{J}_{\theta\lambda})^{-1} \hat{d}_\lambda. \quad (8)$$

여기서 $d_\lambda, J_{\lambda\lambda}, J_{\lambda\theta}, J_{\theta\theta}, J_{\theta\lambda}$ 는 각각 $d_\lambda = \partial l / \partial \lambda = \sum_i \partial l_i(\theta, \lambda) / \partial \lambda$, $J_{\lambda\lambda} = -E[\sum_i \partial^2 l_i / \partial \lambda \partial \lambda^t]$, $J_{\lambda\theta} = -E[\sum_i \partial^2 l_i / \partial \lambda \partial \theta^t]$, $J_{\theta\theta} = -E[\sum_i \partial^2 l_i / \partial \theta \partial \theta^t]$

그리고 $J_{\theta\lambda} = -E[\sum_i \partial^2 l_i / \partial \theta \partial \lambda^t]$ 로 쓰여질 수 있으며, (8)에 쓰인 것은 이 식들에 MLE를 대입한 추정량이며, 이제 (8)에 주어진 스코어검정통계량을 풀면, 다음 결과를 얻는다.

<결과 1> $H_0: \lambda = \lambda_0$ 을 검정하기 위한 스코어검정통계량은 다음과 같이 표현된다.

$$S = \frac{1}{2} V^t \bar{C} [C_\gamma^t C_\gamma - F^t B^{-1} F]^{-1} \bar{C}^t V \quad (9)$$

여기서 V 는 $t \times 1$ 확률벡터로서 다음과 같은 구성요소들을 갖는다.

$$v_i = \sum_{j=1}^{n_i} (e_{ij} - \hat{\xi} \phi_i \bar{e}_i)^2 / \hat{\sigma}^2 - n_i - \hat{\xi} \phi_i \quad (10)$$

위에서, $e_{ij} = y_{ij} - x_{ij}^t \hat{\beta}$, $\bar{e}_i = \sum_{j=1}^{n_i} e_{ij} / n_i$, $\phi_i = n_i / (1 + n_i \hat{\xi})$, 그리고, $\hat{\beta}$, $\hat{\sigma}^2$ 와 $\hat{\xi}$ 은 귀무가설 아래에서 β , σ^2 , ξ , 각각의 최우추정량이다. $w'(\lambda^t z_i) = \partial w(\lambda^t z_i) / \partial \lambda$ 라고 놓자. $w'(\lambda_o^t z_i)$ 는 $\lambda = \lambda_0$ 일때 계산된 값이다. C 는 $t \times q$ 행렬이고 이 행렬의 i 번째 행은 $[w'(\lambda_o^t z_i)]^t$ 이다. C_γ 는 $t \times q$ 행렬이고 i 번째 행은 $\sqrt{\gamma_i} [w'(\lambda_o^t z_i)]^t$ 이다. \bar{C} 는 $t \times q$ 행렬로서 C 의 각 열에서 그 열의 평균을 뺀 값이다. 마지막으로, F^t 와 B 는 각각 $q \times 2$ 그리고 2×2 행렬이고 다음과 같다.

$$F^t = \left(\sum_i \frac{n_i - \hat{\xi} \phi_i}{\hat{\sigma}^2} w'(\lambda_o^t z_i) \quad \sum_i \frac{\phi_i^2}{n_i} w'(\lambda_o^t z_i) \right),$$

$$B = \begin{bmatrix} N/\sigma^4 & \sum_i \phi_i / \hat{\sigma}^2 \\ \sum_i \phi_i / \hat{\sigma}^2 & \sum_i \phi_i^2 \end{bmatrix}$$

<증명> 식 (8)에 의하면, $\hat{d}_\lambda = \frac{1}{2} \bar{C}^t V$, $\hat{J}_{\lambda\lambda} = \frac{1}{2} C_\gamma^t C_\gamma$, $\hat{J}_{\lambda\theta} = \hat{J}_{\theta\lambda} = \frac{1}{2} F^t B^{-1} F$. 스코어통계량을 나타내고 있는 식(9)의 증명은 Ahn(2000)에서 임의효과에 대한 등분산 여부를 검정하는 스코어통계량을 유도하는 과정과 원칙적으로 같은 절차를 밟게 된다. 다만, 조금 더 복잡한 형태를 띠게 되는데 이는 확장모형에서 모수가 임의효과보다 차수가 큰 오차에 붙기 때문이다. 다음 장의 시뮬레이션 경우, Ahn(2000)에서는 설명변수가 하나인 경우만 다루었지만 여기에서는 설명변수가 3개인 경우까지 다루게 될 것이다. 완전한 증명은 상당히 복잡하고 지루한 벡터의 미분을 포함하므로 여기서는 중요한 과정만 보인다. 로그우도함수를 λ 에 대하여 편미분하면 $d_\lambda = \partial l / \partial \lambda = \sum_i \partial l_i(\theta, \lambda) / \partial \lambda$ 로 쓰여진다. $l_i(\theta, \lambda)$ 의 형태는 (7)에서 주어지고 d_λ 의 k 번째 요소는 $d_{\lambda_k} = \sum_i \frac{\partial l_i}{\partial \lambda_k}$ 로서 다음과 같다.

$$-\frac{1}{2} \sum_i [TR(Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k}) + \frac{1}{\sigma^2} \varepsilon_i^t (-Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k} Q_i^{-1}) \varepsilon_i].$$

이 식은 Rogers(1980)에 있는 행렬미분에 대한 아래의 두 결과를 이용하여 얻은 것이다.

$$\frac{\partial}{\partial \lambda_k} \log |Q_i| = TR(Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k}), \quad -\frac{\partial Q_i^{-1}}{\partial \lambda_k} = -Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k} Q_i^{-1}.$$

임의효과에 대한 등분산 여부를 검정하는 스코어통계량을 유도하는 과정과 차이가 나는 것은 위 식에서 Q_i 가 다른 형태를 갖는다는 것이다. 식 (6)을 이용하면 Q_i 는 다음과 같이 표현된다.

$$Q_i = w(\lambda^t z_i) I_i + \xi \mathbf{1}_i \mathbf{1}_i^t$$

그리고 역행렬은 Graybill (1969) 에서와 같이 다음과 같이 표현될 수 있다.

$$Q_i^{-1} = w(\lambda^t z_i)^{-1} (I_i - \xi w(\lambda^t z_i)^{-1} \mathbf{1}_i A^{-1} \mathbf{1}_i^t)$$

$$\text{여기서, } A = \mathbf{1} + \xi w(\lambda^t z_i)^{-1} \mathbf{1}_i^t \mathbf{1}_i.$$

이제 d_{λ_k} 를 구성하고 있는 두 요소는 다음과 같이 쓰여질 수 있다.

$$TR(Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k}) = w'_{ik} (n_i - \xi \phi_i)$$

$$\varepsilon_i^t (Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k} Q_i^{-1}) \varepsilon_i = w'_{ik} \sum_j (\varepsilon_{ij} - \xi \phi_i \bar{\varepsilon}_i)^2$$

이 얻어진다. 여기서 w'_{ik} 는 $\partial w(\lambda^t z_i)/\partial \lambda_k$ 을 $\lambda = \lambda_0$ 로 놓고 구한 것이며, $\bar{\varepsilon}_i$ 는 i 그룹의 평균 즉, $\bar{\varepsilon}_i = \sum_{j=1}^{n_i} \varepsilon_{ij} / n_i$ 이다. 이제, d_{λ_k} 은 다음과 같이 표현된다.

$$d_{\lambda_k} = -\frac{\xi}{2} \sum_i w'_{ik} (n_i - \xi \phi_i - \sum_{j=1}^{n_i} (e_{ij} - \xi \phi_i \bar{e}_i)^2) \quad (11)$$

여기서, $\phi_i = n_i/(1+n_i \xi)$ 이고, w'_{ik} 는 $q \times 1$ 벡터로서 w'_{ik} 을 k 번째 요소로 갖는다. d_{λ_k} 은 $q \times 1$ 벡터인 d_{λ} 의 k 번째 요소로서 추정치 \hat{d}_{λ} 는 다음과 같이 요약된다. 즉, $\hat{d}_{\lambda} = \bar{C}^t V / 2$. 한편 $J_{\lambda\lambda}$ 의 km 번째 요소는 다음과 같이 쓰여진다. $\sum_i w'_{ik} w'_{im} [n_i - \xi \phi_i + \xi \phi_i (\xi \phi_i - 1)] / 2$. $\hat{J}_{\lambda\lambda}$ 은 정의된 C_{γ} 을 이용하여 쓰여진다.

$$\hat{J}_{\lambda\lambda} = \frac{1}{2} C_{\gamma}^t C_{\gamma} \quad (12)$$

$\hat{J}_{\lambda\lambda}$ 을 얻을 때 이용했던 같은 방법을 이용, $\hat{J}_{\lambda\theta}$ 와 $\hat{J}_{\theta\theta}$ 을 구하면 다음 결과를 얻는다.

$$\hat{J}_{\lambda\theta} \hat{J}_{\theta\theta}^{-1} \hat{J}_{\theta\lambda} = \frac{1}{2} \quad F^t \quad B^{-1} F. \quad (13)$$

(11), (12), (13) 을 정리하면 (9) 의 스코어검정통계량을 구한다. (증명 끝)

그룹별 관측치수가 똑같은 균형자료 ($n_i = n$) 의 경우, 스코어검정통계량 S 는 다음과 같이 쓰여진다.

$$S = \frac{1}{2} \ R^t \bar{C} [\bar{C}^t \bar{C}]^{-1} \bar{C}^t R \quad (14)$$

여기서 R 은 $t \times 1$ 벡터이고 구성요소는 $R_i = \phi \bar{e}_i^2 / \hat{\sigma}^2$ 이며, ϕ 는 $\phi = n/(1+n\hat{\xi})$ 이다. 이 경우, S 는 새로 만들어진 모형, $R = \gamma_0 1 + C\gamma + \varepsilon_R$ 에서 R 을 종속변수로 하고 C 를 독립변수로 하여 적합된 회귀모형에서 얻어진 회귀제곱합의 1/2 에 해당한다. 만일 (2)에서 언급된 가중함수 w 로 지수함수를 사용한다면, $a_k = 1$ 일 때 $w'(\lambda_0^t z_i) = z_i$ 이고, $a_k = 0$ 일 때 $w'(\lambda_0^t z_i) = \log(z_i)$ 이 될 것이다. 균형자료에서 얻어진 이 결과는 선형회귀분야에서 Cook 과 Weisberg (1983)에 의해 얻어진 결과와 원칙적으로 동일한 것이다. 즉, 선형회귀모형 $y_i = \beta_0 + x_i^t \beta + \varepsilon_i$ 에서 오차분산의 확장모형으로 $var(\varepsilon_i) = \sigma^2 \exp(\lambda^t z_i)$ 가 사용될 때 $H_0: \lambda = 0$ 을 검정하기 위한 스코어검정통계량은 e_i 가 $e_i = y_i - \hat{\beta}_0 - x_i^t \hat{\beta}$ 으로 표현되고, $\hat{\beta}_0, \hat{\beta}, \hat{\sigma}^2$ 가 귀무가설 아래에서 β_0, β, σ_2 의 최우추정량일 때 $e_i^2 / \hat{\sigma}^2$ 을 종속변수로 하고 z_i 을 독립변수로 하여 적합한 회귀모형에서 얻어진 회귀제곱합의 1/2 에 해당한다는 것이다.

4. 시뮬레이션

귀무가설이 사실일 경우 식(10)의 스코어검정통계량은 χ^2 분포에 근사하다. 표본이 작을 경우에 이러한 χ^2 분포 근사가 얼마나 의미있는지 알아보기 위해서 시뮬레이션이 실시되었다. 스코어검정통계량의 귀무가설 분포에 대한 χ^2 근사가 유한표본을 갖고 있을 경우에 얼마나 정확하고 타당성이 있는 방법인가 알아보기 위해 균형자료와 불균형자료에 대해서 시뮬레이션이 시행되었다. 그룹수 t 는 5, 10, 그리고 20 을 갖도록 하였고, 설명변수의 갯수 p 는 1 과 2 를 갖도록 하였다. 균형자료의 경우, 그룹별 관측치의 갯수 n 은 5 로 고정하였다. 불균형자료의 경우, 그룹별 관측치의 갯수 n_i 는 난수표의 5 번째 열을 골라 아래로 읽으면서 4, 5, 6 에 해당하는 숫자를 골라 20 개를 마련하였다. 골라진 20개의 난수는 5, 6, 5, 5, 4, 4, 6, 6, 4, 6, 6, 4, 4, 5, 5, 4, 6, 5.

4, 4 였다. t 가 5인 경우 이중 처음 5 숫자인 5, 6, 5, 5, 4 를 n_i 로 사용하였다. 공변수 z 는 항상 설명변수 x 와 같도록 하였다. 공변수의 갯수를 q 로 표시하고 있으므로 $q = p$ 가 되도록 하였다. 설명변수행렬, X 로 사용할 행렬을 만들기 위해 표준정규분포에서 임의수를 발생시켜 100×4 의 행렬을 준비하였다. 모형이 절편을 갖도록 하기 위해 X 의 첫번째 열은 모두 1 을 갖도록 하였다. t 와 p 의 값의 조합에 따라 우리는 499 개의 반복적인 표본을 추출하게 되고 각 표본마다 스코어검정통계량을 계산하게 된다. 499 개의 통계량이 모여지면 이것을 가지고 스코어 검정통계량의 분포를 얻게 되는데 이렇게 한개의 스코어검정통계량의 분포를 얻게 되는 절차를 1 개의 시뮬레이션이라 하자. 즉, 우리는 균형인 경우 9개, 그리고 불균형인 경우 9개의 시뮬레이션이 필요하게 되며, 따라서 총 18 개의 시뮬레이션을 하게 된다.

각 시뮬레이션에서 t 와 p 의 값이 결정되면 이 행렬 X 는 필요한 부분을 위에서 준비한 100×4 의 행렬중 왼쪽위 부분을 선택하여 사용하고 각 시뮬레이션이 시행되는 동안에는 값이 변하지 않고 고정될 것이다. 예를 들어 그룹수가 20 이고 설명 변수의 갯수가 2 인 경우를 살펴보자. 균형자료인 경우에는 그룹별 관측치의 갯수가 5 이므로 관측치의 총 갯수는 100 이고 설명변수 2 개에 대한 자료를 얻기 위해 모든 값이 1 인 첫번째 열을 포함하여 총 3 개의 열이 필요하므로 100×4 의 행렬중 왼쪽위 부분을 선택하여 사용하고, 불균형자료의 경우에는 그룹별 관측치 갯수가 5, 6, 5, 5, 4, 4, 6, 6, 4, 6, 6, 4, 4, 5, 5, 4, 6, 5, 4, 4 (총 98개) 이므로, X 의 왼쪽위코너에서 98×3 의 행렬을 취하게 된다. 이제, 그룹수가 10 이고 설명 변수의 갯수가 1 인 경우를 살펴보자. 균형자료인 경우에는 그룹별 관측치의 갯수가 5 이므로 관측치의 총 갯수는 50 이고 설명변수 1 개에 대한 자료를 얻기 위해 모든 값이 1 인 첫번째 열을 포함하여 총 2 개의 열이 필요하므로 X 의 왼쪽위코너에서 50×2 의 행렬을 취하게 된다. 불균형자료의 경우에는 그룹별 관측치 갯수가 5, 6, 5, 5, 4, 4, 6, 6, 4, 6 (총 51개) 이므로, X 의 왼쪽위코너에서 51×2 의 행렬을 취하게 된다. 반응변수 y_{ij} 는 다음 모형에서 발생되었다.

$$y_{ij} = x_{ij}^t \beta + b_i + \varepsilon_{ij}, \quad i=1, \dots, t, \quad j=1, \dots, n.$$

고정효과인 회귀모수 β 는 모두 0 으로 놓았고, σ^2 과 σ_b^2 또한 모두 1 로 놓아졌다. 즉, b_i 와 ε_{ij} 는 모두 기대값이 0 이고 분산이 1 인 표준정규분포에서 임의로 추출되었다.

(표 1)은 t 와 p (또는 q) 의 조합에 따라 행하여진 18 개의 시뮬레이션 결과이다. 네개씩 짹을 지어 세로로 쓰여진 값들은 시뮬레이션에서 얻은 스코어검정통계량의 표본분포에서 읽은 90%, 95%, 97.5%, 그리고 99%에 해당하는 점들이다. 괄호 안의 값은 불균형자료인 경우에 얻어진 결과이다. 예를 들면, $t = 5$ 이고 $q = 1$ 일 때의 (2.54, 3.25, 3.96, 5.69) 과 (2.58, 3.53, 4.78, 7.12) 은 각각 균형자료와 불균형자료에서 얻어진 스코어검정통계량의 표본분포에서 읽은 90%, 95%, 97.5%, 그리고 99%에 해당하는 점들이다. 비교대상이 되는 χ^2 분포 (자유도 $q = 1$)의 % 포인트는 오른쪽 끝 열에 2.71, 3.84, 5.02, 6.63 이 세로로 기재되어 있다.

(표 1) 스코어검정통계량의 귀무가설 분포에 대한 χ^2 근사

p	%점	$t = 5$	$t = 10$	$t = 20$	χ^2
1	0.90	2.54 (2.58)	2.56 (2.88)	2.65 (2.70)	2.71
	0.95	3.25 (3.53)	3.76 (4.25)	4.51 (4.33)	3.84
	0.975	3.96 (4.78)	4.62 (5.66)	6.65 (6.09)	5.02
	0.99	5.69 (7.12)	6.06 (6.25)	7.90 (7.47)	6.63
2	0.90	3.97 (4.12)	4.23 (4.90)	4.76 (4.79)	4.61
	0.95	5.08 (5.31)	5.52 (5.95)	6.10 (6.49)	5.99
	0.975	6.31 (6.98)	6.89 (7.33)	8.14 (8.43)	7.38
	0.99	9.71 (8.67)	8.12 (9.30)	9.93 (10.4)	9.21
3	0.90	5.11 (5.45)	6.16 (6.41)	5.68 (6.33)	6.25
	0.95	6.69 (6.58)	8.08 (8.40)	6.92 (8.37)	7.82
	0.975	7.79 (7.83)	9.35 (10.7)	8.30 (9.60)	9.35
	0.99	8.65 (9.55)	11.2 (12.3)	11.1 (12.9)	11.3

(표 1)에서 보여지듯이 균형자료와 불균형자료 모두, 그룹수 (t) 가 증가할수록 스코어검정통계량의 % 포인트는 χ^2 분포의 해당하는 점에 매우 가까운 것을 보여주고 있으며, 이는 스코어통계량의 점근적인 행태를 뒷받침해 준다고 할 수 있겠다. 그러나, 스코어검정통계량의 % 포인트가 해당 χ^2 값에 비해 일반적으로 작은 것으로 나타나고 있는데 이는 스코어검정을 위해서 χ^2 를 사용할 경우 보수적인 입장으로 취하는 것이라고 할 수 있다. 이는 χ^2 검정을 사용함으로써 실제로 귀무가설을 기각해야하는 경우보다 적게 기각하게 되기 때문이다. 균형자료와 불균형자료로 부터의 결과에서 특이한 사항은 총 36개의 백분위수중 균형자료의 백분위수가 불균형자료의 백분위수보다 큰 경우는 겨우 6개이고 나머지 30개는 반대로 불균형자료의 백분위수가 더 크다. 이 두 경우에 스코어검정통계량의 분포를 비교하여 보면, 불균형자료의 경우에 있어서 꼬리 부분이 더 두터운 형태를 보인다고 할 수 있다.

5. 결론

스코어검정은 통계학의 여러 분야에서 특히 진단 분야에 많이 사용되어 왔다. 이 논문에서도 선형혼합모형에서 오차에 대한 등분산 여부를 검정하는데 유용하게 쓰였다. 여기서 제시된 스코어검정은 등분산 귀무가설이 기각되는 경우 분산이 어느 방향으로 증가하는지를 알아볼 수 있는 근거 또한 제시하고 있다. 스코어검정의 또 다른 장점은 통계량을 계산하기가 매우 용이하다는 것이다. 즉, 귀무가설 아래에서의 최우추정량만으로 검정통계량이 계산되므로 기존의 통계패키지로도 쉽게

구할 수 있기 때문이다. 또한, 접근적으로는 우도비검정과 같을지라도 대립가설 아래에서 추정량을 필요로 하는 우도비검정과 달리 귀무가설 아래에서의 최우추정량만으로 검정통계량이 계산되는 장점을 가지고 있다. 그러나, 일반적으로 우도비검정은 스코어검정보다 더 큰 파워를 갖는다고 알려져 있으므로 이 논문에서 다른 선형혼합모형의 오차에 대한 등분산 문제에 있어서 이 두 검정법에 대한 파워의 비교가 필요하다고 하겠다. 이 논문에서는 불균형자료를 다루고 있는데 불균형의 정도에 따라 결과가 달라질 수 있으므로 시뮬레이션이 4장에서 행하여진 것보다 좀더 다양하고도 광범위하게 실시되어야 한다. 또한, 스코어검정의 그래프적인 방법이 개발되는 것이 중요하다. 검정통계량의 P값과 함께 그래프적인 방법이 제시된다면 더욱 효과적일 것이기 때문이다.

참고문헌

- [1] Ahn, C.H. (2000), "Testing Homogeneity for Random Effects in Linear Mixed Model," *한국 통계학회논문집* 제7권 2호, 403-414.
- [2] Atkinson, A.C. (1985), *Plots, Transformations and Regression*, Oxford : Oxford University Press.
- [3] Bickel, P. (1978), "Using Residuals Robustly I: Tests for Heteroscedasticity, Non-linearity", *Annals of Statistics*, Vol. 6, 266-291.
- [4] Carroll, R.J. 외 Ruppert, D. (1981), "On Robust Tests for Heteroscedasticity," *Annals of Statistics*, Vol. 9, 205-209.
- [5] le Cessie, S 외 van Houwelingen, H.C. (1995), "Testing the Fit of Regression Model via Score Tests in Random Effects Model," *Biometrics*, 51, 600-614.
- [6] Chi, E.M. 과 Reinsel, G.C. (1989), "Models for Longitudinal Data With Random Effects and AR(1) Errors," *Journal of the American Statistical Association*, Vol. 84, 452-459.
- [7] Cook, R.D. (1986), "Assessment of Local Influence (with discussion)," *Journal of Statistical Society, Series B*, 48, 133-169.
- [8] Cook, R.D., Beckman, R. 과 Nachtsheim, C. (1987), "Diagnostics for Mixed-Model Analysis of Variance, *Technometrics* 29, 413-426.
- [9] Cook, R.D. 과 Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman Hall.
- [10] _____ (1983), "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, 70, 1-10.
- [11] Cox, D.R. 과 Hinkley, D.V. (1974), *Theoretical Statistics*, London: Chapman Hall.
- [12] Graybill, F.A. (1969), *Introduction to Matrices with Application in Statistics*, Belmont, California: Wadsworth.
- [13] Hammerstrom, T. (1981), "Asymptotically Optimal Tests for the Heteroscedasticity in the General Linear Model," *Annals of Statistics*, Vol. 9, 368-380.
- [14] Hocking, R. R. (1984), "Diagnostics Methods in Variance Component Estimation," *Proceedings of International Biometrics Conference*, Tokyo, Japan.
- [15] Jacqmin-Gadda 외 Commenges (1995) "Test of homogeneity for generalized linear model," *Journal of the American Statistical Association*, 90, 1237-1246.

- [16] Rogers, G.S. (1980), Matrix Derivatives, Marcel Dekker, New York.
- [17] Serfling (1980), Approximation Theorems of Mathematical Statistics, New York: Wiley.
- [18] Silvey, S.D. (1959), "The Lagrangian Multiplier Test," *The Annals of Mathematical Statistics*, 30, 389–407.
- [19] Smyth, Gordon (1989), "Generalized Linear Models with Varying Dispersion," *Journal of Royal Statistical Society, Series B*, 51, 47–60.
- [20] Verbeke & Lasaffre (1996). "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association*, Vol. 91, 217–221.