The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.361
28(2), 361–369

Analysis of Field Test Data using Robust Linear Mixed-Effects Model

Eun Hee Hong[a] · Youngjo Lee[a] · You Jin Ok[b,1] · Myung Hwan Na[b] · Maengseok Noh[c] · Il Do Ha[c]

[a]Department of Statistics, Seoul National University
[b]Department of Statistics, Chonnam National University
[c]Department of Statistics, Pukyong National University

(Received April 6, 2015; Revised April 9, 2015; Accepted April 9, 2015)

Abstract

A general linear mixed-effects model is often used to analyze repeated measurement experiment data of a continuous response variable. However, a general linear mixed-effects model can give improper analysis results when simultaneously detecting heteroscedasticity and the non-normality of population distribution. To achieve a more robust estimation, we used a heavy-tailed linear mixed-effects model for a more exact and reliable analysis conclusion than a general linear mixed-effects model. We also provide reliability analysis results for further research.

Keywords: Heavy-tailed linear mixed-effects model, linear mixed-effects model, reliability analysis, repeated measurement experiment.

1. 서론(Introduction)

타이어 수명 실험에서와 같이 타이어를 차량에 부착한 후 주행거리에 따라 타이어의 마모상태를 측정하는 경우에는 하나의 동일한 타이어에 대해서 관심 있는 반응변수를 여러 번 반복하여 측정한 반복측정 자료(repeated measured data) 형태로 관측된다. 반응변수가 연속형일 때는 반복측정 자료를 분석하기 위해서는 대부분 선형혼합모형(linear mixed models)의 사용이 제안되고 있다. 선형혼합모형에서 변량효과와 요인의 오차는 일반적으로 정규분포로 가정한다. 그러나, 정규성의 가정이 의심이 되는 경우, 예를 들면 자료가 정규분포에 맞지 않는 왜도, 두터운 꼬리, 이상치 등의 일반적이지 않은 특성을 가지는 경우에 선형혼합모형은 편의된 추정치를 제시하게 되며, 이로 인한 적절치 않은 가설검정 결과

This work was supported by the Nuclear Safety Research Program through the Korea Radiation Safety Foundation(KORSAFe) and the Nuclear Safety and Security Commission(NSSC), Republic of Korea (Grant No. 1305033).
This research was supported by an NRF grant funded by Korea government (MSIP) (No. 2011-0030810).
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology, Korea (No. 2010-0021165).
[1]Corresponding author: Department of Statistics, Chonnam National University, Gwangju 500-757, Korea. E-mail: okyoujin@hanmail.net

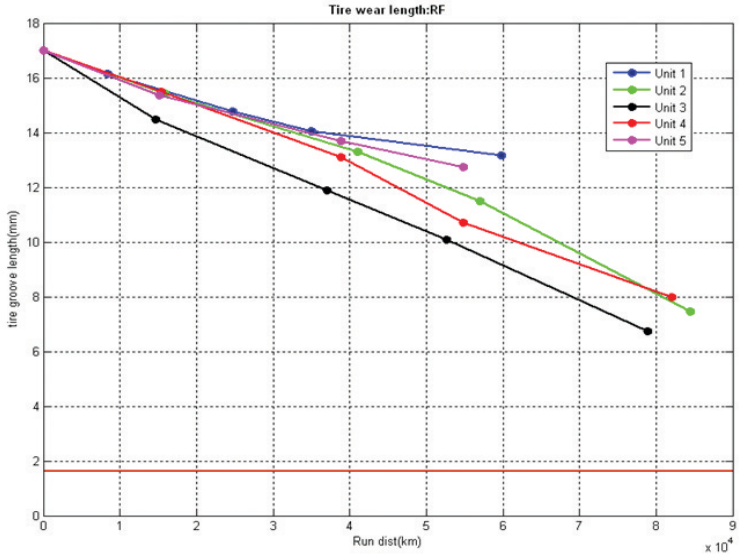

Figure 2.1. The plot about the wear level of right side tire at the front of cars. The vertical line corresponds tire groove depth (mm) on each acceleration experiment. The horizon line corresponds the mileage (km) of cars.

를 주게 된다. 이러한 문제점을 해결하기 위해 Lange와 Shinsheimer (1993)는 선형혼합모형의 로버스트 추론을 함으로써 비대칭 정규/독립 분포(Skew-Normal/Independent; SNI)를 사용할 수 있음을 제시하였다. Branco와 Dey (2001)는 다변량의 기울어진 정규/독립 분포를 제안하였다. SNI 분포는 특이한 경우들로서 기울어진 t-분포(skew-t), 기울어지고 끊어진(skew-slash) 정규분포, 기울어지고 오염된(skew-contaminated) 정규분포를 포함한다.

Wakefield 등 (1994)는 베이지안에서 변량효과에 대한 t분포를 유도하였다. 그러나 방법론적으로 변량효과모형에 두터운 꼬리 분포를 사용하였을 때의 모수통계량이 적절한지에 대해서는 회의적이다. Lee와 Nelder (2006)는 반응변수의 두터운 꼬리 분포를 모델링하는 체계적인 방법을 소개하였고 이 모형은 로버스트 통계량과 추론을 제공하고 있다 (Lee와 Noh, 2012).

본 논문은 로버스트 통계를 이용한 두터운 꼬리 분포를 가진 선형혼합모형을 실제 필드 데이터에 적용하였다. 2절에서는 데이터 분석을 제시하였고, 3절에서 모델을 적합하는 과정과 결과를 제시하였다. 선형혼합모형에서 집단과 요인간의 이분산성을 해결하는 좋은 방법과 변량효과의 유용성을 언급한다. 4절에서는 신뢰성분석 결과를 제공한다. 5절에서는 제기된 문제에 대한 논의와 결론을 기술한다. 모든 결론과 모형 검증 및 도표는 R 패키지 Noh와 Lee (2011)가 개발한 dhglm을 사용하였다.

2. 타이어 필드 시험 자료(Field Experiment Data of Tire)

분석에 사용한 데이터는 한 종류의 타이어에 대해 필드시험 후 측정한 관측값이다. 다섯 대의 서로 다른 차량에 대해 동일 타이어를 장착하여 운행하였으며 시험자는 4개월 동안 4회 시험하면서 4회 반복 측정하였다. 타이어는 다섯 종류의 차량의 양쪽 앞바퀴에 각각 장착되었다. 매 시험마다 주행을 한 후에 주행거리를 측정하였다. 각 차량의 양쪽 앞바퀴에 대해서 각각 관측을 하였다. 총 자료의 관측수는 총 40회 (= 5(차량의 종류) * 4(반복측정횟수) * 2(오른쪽, 왼쪽 바퀴))이며 한 차종에서 반복측정의 마지막 1회는 결측치이다. 이 자료를 선형혼합모형에 적합하여 모수값을 추정하였고 잔차는 두터운 꼬리 분포

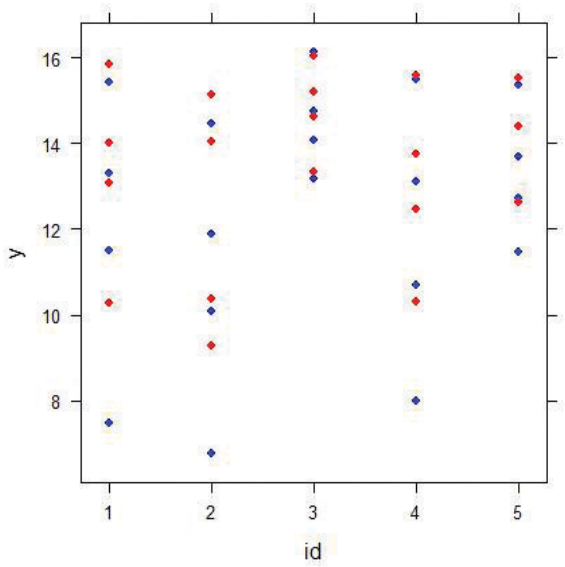

Figure 2.2. The plot describing the tire wear data: The blue point corresponds the left front tire and the red point corresponds the right front tire. The tire groove depth of fifth car on the last experiment is missing.

를 가졌다. 적합된 모형의 추정값으로부터 타이어의 수명 기대치를 예측하여 4절에서 제시하고 있다.

자료는 총 4개의 변수로 구성된다. 'id' 변수는 차의 종류이다. 'location' 변수는 차량의 앞바퀴의 위치를 가리키며 오른쪽과 왼쪽으로 구분된다. 'id' 변수와 'location' 변수는 반응변수에 대해서 범주형 설명변수들이다. 'mileage'는 주행거리이고 'depth'는 타이어 홈의 깊이이며 반응변수이다. 주행거리는 차량에 장착된 계기판으로부터 측정되었고 홈의 깊이는 타이어의 Grooves 부분이 측정되었다. 'mileage' 변수와 'depth' 변수는 각 시험마다 측정된 값으로 연속형이고 양수이며 단위는 각각 km와 mm이다. 'depth' 변수의 초기값은 모든 차종에서 17mm로 동일하며 시험이 진행될수록 Figure 2.1과 같이 점차 감소한다.

Figure 2.2에서 점들을 수직적으로 보면 각 차종에서의 홈의 깊이를 알 수 있고 수평적으로 보면 차종간 평균 차이를 알 수 있다. 푸른색 점과 붉은색 점은 오른쪽, 왼쪽 바퀴를 의미한다. 차종간 타이어 홈의 깊이의 차이가 매우 명확하게 나타난다. 이것은 분석자가 차종에 있어서 선형혼합모형의 변량효과를 고려해야함을 말한다. 변량효과에 대한 추정값은 Verbeke와 Molenberghs (2003)가 제안한 50대 50 혼합 카이제곱 분포(mixture χ^2-distribution)를 사용한 변량효과에 대한 REML 기반 우도비 검정을 3절에서 제시하고 있다 (Self와 Liang, 1987; Ha 등, 2012).

3. 데이터 분석(Data Analysis)

3.1. 모형 적합(Model fitting procedure)

2절에서 제시된 데이터에 대한 선형혼합모형은 다음과 같다.

$$\begin{aligned} \text{depth}_{ij} &= \beta_0 + \beta_1 \text{mileage}_{ij} + \beta_2 I(\text{location} = \text{right})_{ij} + v_i + e_{ij}, \\ & v_i \sim N\left(0, \sigma_v^2\right),\ e_{ij} \sim N\left(0, \sigma_\epsilon^2\right),\ i = 1, \ldots, 5,\ j = 1, \ldots, 4. \end{aligned} \tag{3.1}$$

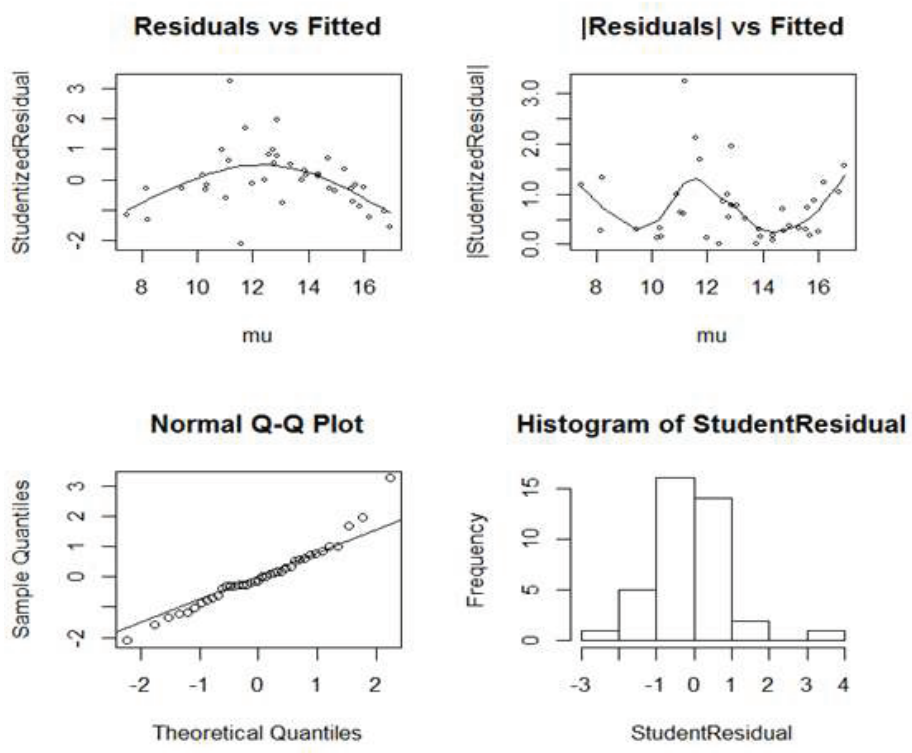

Figure 3.1. The panel on the top left side in the figure is the plot of studentized residuals versus the fitted means. The left side panel at the bottom is the normal Q-Q plot which provides a graphical way to determine the level of normality.

단, $I(\cdot)$는 오른쪽 타이어일 때를 가리키는 지시함수이다. 반응변수는 각 실험에서 측정한 홈의 깊이이다. 'mileage' 변수와 'location' 변수는 서로 독립인 설명변수이다. 여기서 v_i는 차종에 따른 변량효과이고 e_{ij}는 오차항이다. v_i와 e_{ij}는 서로 독립이며 정규분포로 가정한다. 모형 체크 도표가 Figure 3.1에 나와 있다. 적합된 모형은 등분산성을 따르지 않고 있다. 등분산성을 따르게 하기 위하여 여러 가지 변환을 하였다. 가장 잘 적합된 모형은 반응변수의 제곱변환이고 식 (3.2)와 같다.

$$\begin{aligned} \text{depth}_{ij}^2 &= \beta_0 + \beta_1 \text{mileage}_{ij} + \beta_2 I(\text{location} = \text{right})_{ij} + v_i + e_{ij}, \\ v_i &\sim N\left(0, \sigma_v^2\right),\ e_{ij} \sim N\left(0, \sigma_\epsilon^2\right),\ i = 1, \ldots, 5,\ j = 1, \ldots, 4. \end{aligned} \tag{3.2}$$

모형 (3.2)의 변량효과에 대해 적합된 잔차의 분포와 도표는 Figure 3.2와 같다. 잔차는 등분산성을 잘 따르는 것으로 보인다. 그러나 Q-Q 도표를 보면 비정규성을 가짐을 볼 수 있다. 분포의 양쪽 끝에서 이상치가 보인다. 이것은 변형된 선형혼합모형의 적합검정이 최강검정이 아닐 수 있고 적절한 결과가 도출되지 않을 수 있음을 말한다.

Lee와 Nelder (2006)는 잔차에 있어서 두터운 꼬리를 가진 분포를 다양하게 제안하였다. Noh와 Lee (2007)는 오차변동에서의 변량효과를 제안하였으며 결과통계량은 이상치에 대해 강하다. 또한, 분산모수를 추정하면 결과통계량이 정규분포를 따른다. 분산성분에 대해 변량효과를 고려하는 것은 왜도와 첨

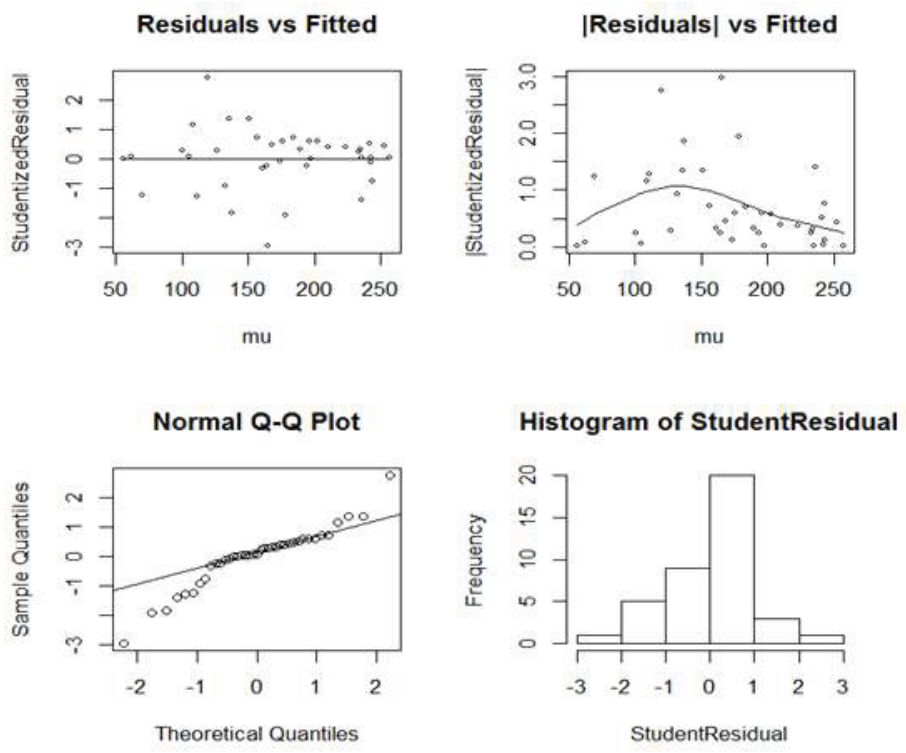

Figure 3.2. The model-checking plot of the transformed model (3.2) for homoscedasticity.

도의 모형식을 세울 수 있게 하여 모형식을 적합한 경우에 분포가정에 덜 민감한 통계량을 제공한다 (Lee 등, 2006). 그러므로 다음과 같은 두터운 꼬리를 가진 선형혼합모형으로서 로버스트 통계량의 추정 및 검정을 생각할 수 있다.

$$\begin{aligned} \text{depth}_{ij}^2 &= \beta_0 + \beta_1 \text{mileage}_{ij} + \beta_2 I(\text{location} = \text{right})_{ij} + v_i + e_{ij}, \\ & e_{ij} \sim \sigma_{ij} z_{ij}, \quad z_{ij} \sim N(0,1),\ \phi_{ij} = \sigma_{ij}^2, \\ & \log \phi_{ij} = \gamma + b_i, \quad b_i \sim N(0,\tau), \\ & i = 1,\ldots,5,\ j = 1,\ldots,4. \end{aligned} \tag{3.3}$$

단, e_{ij} 또는 y_{ij}의 첨도는 ϕ_{ij}가 상수값일 때 $E(\phi_{ij}^2)/E(\phi_{ij})^2 \geq 3$가 된다. 분산의 변량효과 b_i를 고려함으로써 차종간 분산의 변량효과를 표현할 수 있게 되었다. 이 값들은 측정단위에 따라 급격한 변화를 가진다. 또한 변량이 $a_i = \exp(b_i)k/\chi_k^2$과 같다면 오차항 $e_i = (e_{i1},\ldots,e_{in})^t$는 다변량 t분포를 따른다 (Lange 등, 1989). k가 1의 값을 가지면 Cauchy 분포를 따른다. 따라서 분산 성분의 변량효과를 고려하면 다양한 두터운 꼬리 선형혼합모형을 얻을 수 있다. 식 (3.3)을 적합하면 e_{ij}는 자유도 4인 t분포로 볼 수 있다. 적합된 모형의 모형체크 도표가 Figure 3.3에 나와 있다. 모형체크 도표는 적합된 모형의 가정이 잘 만족함을 보여주며, 표준정규 선형혼합모형 (3.1)과 변형된 선형혼합모형 (3.2)와 비교해서도 가정이 잘 만족한다. Table 3.1은 모형 (3.3)의 추정 통계량이다. Table 3.1로부터 모든 공변량이 통계적으로 매우 유의함을 알 수 있다.

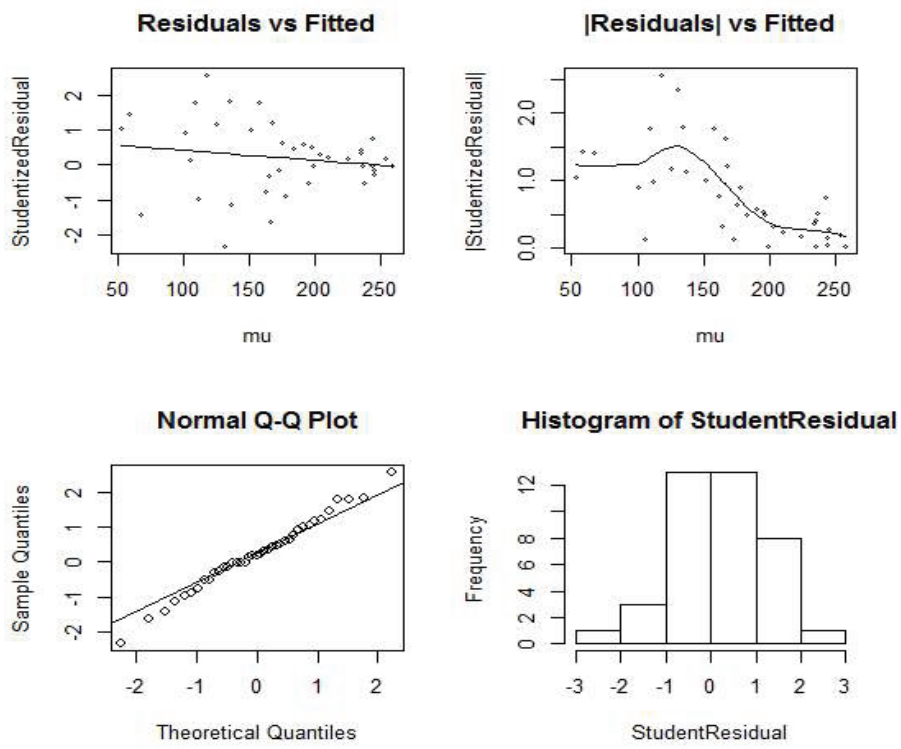

Figure 3.3. The residual plots of heavy-tailed linear mixed-effects model (3.3) and (3.4).

Table 3.1. The estimation results of models (3.3) and (3.4).

Variable	Estimate	Std. Error	t-value	p-value
intercept	279.234	4.8280	57.84	<0.0001
mileage	−20.632	0.7646	−26.99	<0.0001
$I(\text{location} = \text{right})$	−5.796	4.3320	−13.38	<0.0001
$\log(\sigma_v^2)$	−2.180	0.6633		
γ	0.0221	0.2237		
$\log \tau$	−0.7975	0.5106		

3.2. 변량효과의 유의성검정(Testing significance of random effects)

이 절에서는 차종간 변량효과를 포함하는 것이 모형 (3.3)에서 적절한지에 관하여 검증한다. 즉, 차종간 변량효과가 유의한지에 대해 가설을 검증한다. 그 가설은 다음과 같다.

$$H_0 : \sigma_v^2 = 0, \qquad H_A : \sigma_v^2 > 0. \tag{3.4}$$

귀무가설이 의미하는 것은 "차종간 변량효과의 분산이 없다"이다. 이에 반해서 대립가설은 "차종간 변량효과의 분산이 있다"이다. 차종간 변량효과의 분산이 유의하다면 차종간의 변량효과를 선형혼합모형에서 고려해야만 한다는 것을 의미한다.

가설 (3.4)의 검증을 위하여 REML-based 우도비검정(REML-based likelihood ratio test)을 사용하

Table 4.1. The fitted random effects.

Variable	Estimate
1	5.5494
2	−26.6156
3	15.6829
4	−3.1972
5	8.5805

Table 4.2. The predicted life time of both sides of a tire at the front of the cars. the unit is km.

id	left	right	left-right
1	225,035	177,873	47,162
2	209,445	165,702	43,743
3	229,946	181,707	48,239
4	220,795	174,563	46,232
5	226,504	179,020	47,484

고 있다. 검정통계량은 차종간 변량효과를 고려한 모형과 차종간 변량효과를 고려하지 않은 모형의 −2REML 로그우도값의 차이값이다. 이 차이값은 22.7로 계산되어진다. 차종간 변량효과의 분산의 귀무가설이 모수공간의 경계 0에 있으므로 점근적 귀무가설하에서 검정통계량의 값은 χ^2_0과 χ^2_1의 중간값, 즉 가중치를 0.5로 한 값이다 (Verbeke와 Molenberghs, 2003). 검정의 유의성을 알아보기 위해 p-value를 계산하면 다음과 같다:

$$p\text{-value} = 0.5 \times P\left(\chi^2_0 \geq 22.7\right) + 0.5 \times P\left(\chi^2_1 \geq 22.7\right) = 0.5 \times P\left(\chi^2_1 \geq 22.7\right) < .0001.$$

이로부터 귀무가설은 기각되어 차종간 분산의 변량효과는 유의하다. 따라서, 두터운 꼬리를 갖는 로버스트 선형혼합모형 (3.3)에서 차종간 변량효과는 매우 유의함을 알 수 있다.

4. 신뢰성 분석(Reliability Analysis)

이 절에서는 추가적으로 신뢰성 분석 결과를 제시한다. 일반적으로 신뢰성 분석의 주된 목적은 제품의 수명시간을 예측하는 것이라고 알려져 있다. 본 논문은 변량효과를 고려하여 자동차 앞바퀴의 양쪽 타이어의 수명시간을 예측하였다. 왜냐하면 변량효과를 고려한 모형 (3.3)이 통계적으로 매우 유의하기 때문이다. 일반적으로 자동차 타이어의 수명은 타이어의 홈의 깊이가 2.0mm일 때로 알려져 있다. 적합된 모형 (3.3)에서 반응변수 'depth'(홈의 길이)가 2.0mm일 때, 각각의 차종에서의 'mileage'(주행거리)의 추정치는 예측되는 수명시간이 된다. Table 4.1과 Table 4.2는 각각 차종별 적합된 변량효과와 예측된 수명시간을 나타낸다. Table 4.2에서 세 번째 차종에서 왼쪽과 오른쪽 타이어의 수명시간의 차이가 가장 크며 두 번째 차종에서 왼쪽과 오른쪽 타이어의 수명시간의 차이가 가장 작음을 볼 수 있다. 이러한 결과는 Table 4.1의 차종별 변량효과 추정치가 반영되었음을 알 수 있다 (Paik 등, 2015).

5. 결론 및 토의(Conclusion and Discussion)

Noh와 Lee (2011)의 R 패키지 dhglm의 모형체크도표를 사용함으로써 가장 좋은 선형혼합모형을 알아내었다. 모형체크도표는 그래픽적으로 쉽게 모형의 가정을 평가할 수 있게 해준다. 모집단 분포의 등분산성과 비정규성을 찾아냄과 동시에, 반응변수의 제곱변환과 분산의 변량효과를 포함시킴으로써 오차항의 가정을 잘 만족하는 선형혼합모형을 찾을 수 있다. 결론적으로 두터운 꼬리를 허용하는 선형혼합모

형은 이상치에 덜 민감한 통계량을 제공한다. 적합된 모형은 통계적으로도 적절한 추정과 검정을 가능하게 한다. 또한, 우리는 타이어의 수명 기대치가 두터운 꼬리 선형혼합모형을 사용한 변량효과를 따름을 예측할 수 있었다. 예측의 결과로서, 세 번째 차종은 양쪽 타이어의 주행거리 차이에서 가장 큰 차이를 보였고 두 번째 차종에서는 가장 작은 차이를 보였다.

References

Branco M. and Dey D. (2001). A general class of multivariate skew-elliptical distribution, *Journal of Multivariate Analysis*, **79**, 93–113.

Ha, I. D., Noh, M. and Lee, Y. (2012). frailtyHL: A Package for fitting frailty models with h-likelihood, *R Journal*, **4**, 28–37.

Lange, K. and Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression, *Journal of Computational and Graphical Statistics*, **2**, 175–198.

Lange, K., Little, J. A. and Taylor, M. G. J. (1989). Robust statistical modeling using the t distribution, *Journal of American Statistical Association*, **84**, 881–896.

Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion), *Applied Statistics*, **55**, 139–185.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Chapman and Hall, London.

Lee, Y. and Noh, M. (2012). Modeling random effect variance with double hierarchical generalized linear models, *Statistical Modelling*, **12**, 487–502.

Noh, M. and Lee, Y. (2007). Robust modeling for inference from GLM classes, *Journal of American Statistical Association*, **102**, 1059–1072.

Noh, M. and Lee, Y. (2011). dhglm: Double hierarchical generalized linear models. R package version 1.0, Available at `http://CRAN.R-project.org/package=dhglm`

Paik, M. C., Lee, Y. and Ha, I. D. (2015). Frequentist inference on random effects based on summarizability, *Statistica Sinica*, In press.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association*, **82**, 605–610.

Verbeke, G. and Molenberghs, G. (2003). *Repeated Measures and Multilevel Modelling*, Oxford, Eolss.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1994). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler, *Applied Statistics*, **43**, 201–222.

로버스트 선형혼합모형을 이용한 필드시험 데이터 분석

홍은희[a] · 이영조[a] · 옥유진[b,1] · 나명환[b] · 노맹석[c] · 하일도[c]

[a]서울대학교 통계학과, [b]전남대학교 통계학과, [c]부경대학교 통계학과

(2015년 4월 6일 접수, 2015년 4월 9일 수정, 2015년 4월 9일 채택)

요 약

연속측도의 반응변수가 반복측정된 실험 자료의 분석을 위해 흔히 선형혼합모형이 사용된다. 그러나, 잔차의 분포가 이분산성이거나 비정규성을 가질 때 표준적인 선형혼합모형은 적절하지 않은 결과를 가져온다. 잔차의 분포가 두터운 꼬리를 가진 비정규분포를 보이는 타이어 필드시험 데이터를 로버스트 선형혼합모형에 적합시킴으로써 보다 더 정확하고 신뢰할 수 있는 분석결과를 얻을 수 있다. 추가적으로 신뢰성 분석 결과를 제시한다.

주요용어: 두터운 꼬리 분포, 로버스트 통계량, 변량효과, 선형혼합모형, 신뢰성 분석.

본 연구는 원자력안전위원회와 한국방사선안전재단의 지원을 받아 수행한 원자력안전연구사업의 연구결과입니다 (No. 1305033).

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2011-0030811).

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (No. 2010-0021165))

[1]교신저자: (500-757) 광주광역시 북구 용봉로 77, 전남대학교 통계학과. E-mail: okyoujin@hanmail.net