An efficient resampling scheme for outlier detection in linear mixed models

Lu Wang

Supervisors:

Dr. Alison Smith Prof. Brian Cullis Dr. Carole Birrell

AASC 2024

Mixed Models and Experimental Design Lab (MMaED Lab)
National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong
luw@uow.edu.au

Outline

Overview

Model fitting

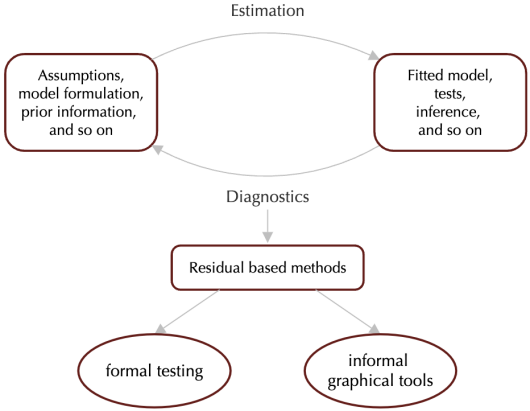


Figure 1: Schematic outline of model fitting. Adapted from Box (1979, 1980) and Cook and Weisberg (1982).

Diagnostics for outlier detection

- Cook et al. (1982) proposed an alternative outlier model (AOM) for ordinary linear models and used maximum likelihood estimation.
  - The AOM model can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{e}_i \\ \text{where } \boldsymbol{e}_i &= \boldsymbol{e} + \boldsymbol{d}_i\delta_i \end{aligned}$$

  - $\boldsymbol{d}_i$ is an $n \times 1$ vector with one in position i and zeros elsewhere;
  - $\text{var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$ and $\delta_i \sim N(0, \omega_i\sigma^2)$.
  - In this model, the variance of all observations apart from the ith is assumed to be σ^2 and the ith has variance $(1 + \omega_i)\sigma^2$, where $\omega_i > 0$.

Overview

Diagnostics for outlier detection

- Cook et al. (1982) proposed an alternative outlier model (AOM) for ordinary linear models and used maximum likelihood estimation.
  - The AOM model can be written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{e}_i \\ \text{where } \boldsymbol{e}_i &= \boldsymbol{e} + \boldsymbol{d}_i\delta_i \end{aligned}$$

  - $\boldsymbol{d}_i$ is an $n \times 1$ vector with one in position i and zeros elsewhere;
  - $\text{var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$ and $\delta_i \sim N(0, \omega_i\sigma^2)$.
  - In this model, the variance of all observations apart from the ith is assumed to be σ^2 and the ith has variance $(1 + \omega_i)\sigma^2$, where $\omega_i > 0$.
- Thompson (1985) considered the same model but used residual maximum likelihood (REML) for estimation.

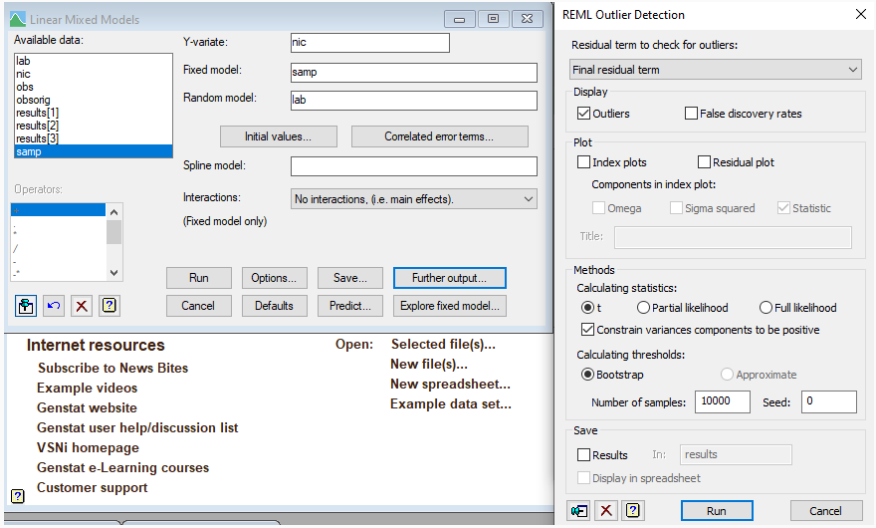
Diagnostics for outlier detection

- Gumedze et al. (2010) extended the AOMs of Cook et al. (1982) to one class of linear mixed models , i.e., variance component models and used REML estimation;
  - a variance shift outlier model (VSOM);
  - assumes independent random effects (including errors);
  - uses score test and offers a threshold for multiple testing;
  - has been implemented in Genstat.

Nicotine example in Genstat

- 138 data points from 14 labs, 10 samples/lab (2 were missing).
- Model:
  - response: nicotine content;
  - fixed: 1/sample;
  - random: lab;
  - independent errors.

Nicotine example in Genstat



Nicotine example in Genstat

Variance shift outlier model

Analysis for residual term

Outlier detection based on test statistic t^2

Thresholds based on bootstrap with 10000 simulated data sets

Units above test-wise threshold 0.0001 < p <= 0.001

Unit	Omega	Residual variance	Test statistic
117	18.07	0.0006891	14.39
31	17.94	0.0006892	14.37
118	16.58	0.0006951	13.39
138	14.70	0.0007029	12.11

Units above test-wise threshold 0.001 < p <= 0.01

Unit	Omega	Residual variance	Test statistic
130	11.112	0.0007179	9.628
129	9.740	0.0007240	8.627
137	8.686	0.0007290	7.808

Units above test-wise threshold 0.01 < p <= 0.05

Unit	Omega	Residual variance	Test statistic
9	4.183	0.0007498	4.373
109	3.969	0.0007507	4.223

Units above experiment-wise threshold (p=0.05) on order statistics

Unit	Omega	Residual variance	Test statistic	Threshold
117	18.07	0.0006891	14.39	12.196
31	17.94	0.0006892	14.37	8.637
118	16.58	0.0006951	13.39	7.180
138	14.70	0.0007029	12.11	6.240
130	11.11	0.0007179	9.63	5.587
129	9.74	0.0007240	8.63	5.098
137	8.69	0.0007290	7.81	4.716

Motivation and future work

Table 1: Type I errors ($\alpha = 0.05$) of score test statistics for a VSOM model in a one-way random effects ANOVA with p groups and r replicates per group, variance ratio γ and residual variance $\sigma^2 = 1$. Threshold values calculated from empirical distribution. 500 data sets were generated for each parameter combination, with percentiles of the empirical distribution of the test statistics calculated from 2500 simulations per data set under the null hypothesis (Gumedze et al., 2010).

p	r	γ	empirical distribution	p	r	γ	empirical distribution
12	3	0.1	0.052	24	3	0.1	0.056
		1	**0.042**			1	**0.060**
		10	0.046			10	**0.042**
6	6	0.1	0.050	12	6	0.1	0.050
		1	**0.036**			1	0.054
		10	**0.040**			10	0.048
3	12	0.1	0.052	6	12	0.1	**0.066**
		1	0.048			1	**0.060**
		10	**0.064**			10	0.054
				3	24	0.1	**0.038**
						1	**0.062**
						10	**0.034**

Motivation and future work

- Focus on score test for linear mixed models with more complex variance structures for errors and random effects,
  - correlated errors - separable autoregressive models for spatial variation in field trials;
  - correlated random effects - inclusion of genetic relatedness via either pedigree or marker data.
- Address multiple comparison issues
  - computational efficiency - quick parametric bootstrap for thresholds;
  - full assessment of experiment-wise type I error rates.
- Implemented in DWReml (David Butler, pers comm).

References i

R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. New York: Chapman and Hall, 1982.

R. D. Cook, N. Holschuh, and S. Weisberg. A note on an alternative outlier model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):370–376, 12 1982. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1982.tb01215.x. URL https://doi.org/10.1111/j.2517-6161.1982.tb01215.x.

F.N. Gumedze, S. Welham, B Gogel, and R Thompson. A variance shift model for detection of outliers in the linear mixed model. *Computational Statistics and Data Analysis*, 54(9): 2128–2144, 2010.

R. Thompson. A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1):53–55, 1985.