Confidence Region in Linear Mixed Model for Longitudinal Data

G. Wimmer

Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovak Republic
Email: wimmerg@mat.savba.sk

Abstract. *In many situations, the response variable is observed on subjects at several time points. Such a data are often referred to as longitudinal data. A widespread model for analysing this type of data is a linear mixed model. We use a score algorithm to estimate a parameters of this model with AR(1) errors to achieve confidence regions for regression parameters. In simulation study we discus qualities of constructed confidence regions and point out some of their deficiencies.*

Keywords: Confidence Region, Linear Mixed Model, Longitudinal Data

1. Introduction

The main characteristic of longitudinal studies is that subjects are measured in some time points or time intervals. With this repeated observations on several subjects we try to describe a common feature, which defines the behaviour of all subjects in time. It is natural to assume that these vectors of outcomes are independent between subjects, but the repeated measurements, done on the single subject exhibit some form of correlation. This is due to the fact that every subject has except these joint attributes also his own individual effects, which affects the final outcome of his repeated measurements. For analyzing such type of data it appears to be advantageous to use the linear mixed model as in [8], which reflects both (common and individual) effects of each subject on his repeated measurements.

It is advisable to note, that individual effects can differ from subject to subject and so the suggested model is able to distinguish between them. Moreover, such model can be used to make statistical inferences not only about the common effects, but also in a broader problem to estimating MSE of the prediction error of the individual's effects, since we are able to estimate with linear mixed model these individual effects and evaluate a deviation of each subject from a common mean. However, this is not the problem discussed in this article.

Apart from that, there is one more advantage to linear mixed model. It permits a certain form of dependence connected with random errors.

2. Model construction

Let us consider a linear mixed model as in [1] with AR(1) errors. The response vector for i-th subject ($i = 1, 2, \ldots, I$) can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\mathbf{X}_i$ is $(n_i \times p)$-dimensional known matrix for i-th subject, $\boldsymbol{\alpha}$ is p-dimensional vector of the unknown regression parameters. These are identical for all subjects. $\mathbf{Z}_i$ are $(n_i \times q)$-dimensional known design matrices for individual effects $\mathbf{b}_i$, where $\mathbf{b}_i$'s are q-dimensional random vectors from $N(\mathbf{0}, \mathbf{D})$ mutually independent. $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$ are n_i-dimensional error vectors independent of $\mathbf{b}_i$. Here $\mathbf{D}$ and $\mathbf{R}_i$ are some covariance matrices. In light of the above mentioned assumptions we require, that for i-th subject ($i = 1, 2, \ldots, I$) at given time point j ($j = 1, 2, \ldots, n_i$) is

$$\varepsilon_{i,j} = \rho\varepsilon_{i,j-1} + \tau_{i,j} \tag{2}$$

where $\tau_{i,j} \sim N(0, \sigma^2)$. ρ is coefficient of autoregression and σ^2 is some positive scalar.

Covariance matrix for random vector $\mathbf{y}_i$ is then

$$Var(\mathbf{y}_i) = \mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^{'} + \sigma^2 \mathbf{R}_i \tag{3}$$

where

$$\sigma^2 \mathbf{R}_i = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}.$$

Let us denote $\boldsymbol{\nu} = (d_{11}, d_{12}, \dots, d_{22}, \dots, d_{rr}, \sigma^2, \rho)'$ as a vector of all variance-covariance parameters in model (1) (i.e. we can write $Var(\mathbf{y}_i) = \mathbf{V}_i(\boldsymbol{\nu})$). Now it is clear that all the parameters of the proposed model (1) are $(\boldsymbol{\alpha}', \boldsymbol{\nu}')'$.

3. Parameter estimation and its properties

With assumption form the section 2 we can use a score algorithm to estimate the parameters of model (1) directly from the likelihood function. Despite the fact, that in our primary interest is to estimate unknown regression parameter $\boldsymbol{\alpha}$, it is necessary to estimate also the variance-covariance parameters $\boldsymbol{\nu}$, since these are usually also unknown. To obtain an estimator of these unknown parameters $(\boldsymbol{\alpha}', \boldsymbol{\nu}')'$, we can use a logarithm of likelihood function, which is proportional to

$$l = -\frac{1}{2} \sum_{i=1}^{I} ln \left| \mathbf{V}_i(\boldsymbol{\nu}) \right| - \frac{1}{2} \sum_{i=1}^{I} \left[(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\alpha})^{'} \mathbf{V}_i^{-1}(\boldsymbol{\nu}) (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\alpha}) \right] \tag{4}$$

With a given maximum likelihood estimate of $\boldsymbol{\nu}$, $\hat{\boldsymbol{\nu}}$, maximizing (4) we get the following maximum likelihood estimate of $\boldsymbol{\alpha}$

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^{I} \mathbf{X}_i^{'} \mathbf{V}_i^{-1}(\hat{\boldsymbol{\nu}}) \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{I} \mathbf{X}_i^{'} \mathbf{V}_i^{-1}(\hat{\boldsymbol{\nu}}) \mathbf{y}_i \right) \tag{5}$$

As we can see, it is necessary at first to estimate unknown variance-covariance parameters and then we can calculate an estimate for regression parameter. Successive iteration between these two steps yields maximum likelihood estimates of unknown parameters, such as $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\nu}}$. For more details see [1].

With a given estimate $\hat{\boldsymbol{\alpha}}$ we can consider a construction of confidence regions for some linear combination $\mathbf{L}\boldsymbol{\alpha}$, where $\mathbf{L}$ is known $(r \times p)$-dimensional matrix. If a covariance matrix of vectors $\mathbf{y}_1, \dots, \mathbf{y}_I$ were known, then $\hat{\boldsymbol{\alpha}}$ is asymptotically normally distributed with mean

$$E(\hat{\boldsymbol{\alpha}}) = \boldsymbol{\alpha} \tag{6}$$

and covariance matrix

$$Var(\hat{\boldsymbol{\alpha}}) = \left(\sum_{i=1}^{I} \mathbf{X}_i^{'} \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \tag{7}$$

In this case is

$$X^2 = (\mathbf{L}\hat{\boldsymbol{\alpha}} - \mathbf{L}\boldsymbol{\alpha})^{'} (\mathbf{L}\, Var(\hat{\boldsymbol{\alpha}}) \mathbf{L}^{'})^{-1} (\mathbf{L}\hat{\boldsymbol{\alpha}} - \mathbf{L}\boldsymbol{\alpha}) \tag{8}$$

χ^2 distributed with r degrees of freedom.

If, and it is in the majority of practical application, must be estimated also the variance-covariance parameters of the model, we can replace (7) with its maximum likelihood estimate

$$\hat{Var}(\hat{\boldsymbol{\alpha}}) = \left(\sum_{i=1}^{I} \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i\right)^{-1},$$

where $\hat{\mathbf{V}}_i = \mathbf{V}_i(\hat{\boldsymbol{\nu}})$ is maximum likelihood estimate of the covariance matrix. Then we can "naïve" assume (and it is that, what many authors do, see for example [2] or [3]) that

$$X^2 = (\mathbf{L}\hat{\boldsymbol{\alpha}} - \mathbf{L}\boldsymbol{\alpha})'(\mathbf{L}\,\hat{Var}(\hat{\boldsymbol{\alpha}})\mathbf{L}')^{-1}(\mathbf{L}\hat{\boldsymbol{\alpha}} - \mathbf{L}\boldsymbol{\alpha}) \tag{9}$$

has also χ^2 distribution with r degrees of freedom.

It was created a MATLAB algorithm "CONFZON" which evaluates the 95% confidence region from (9) for different numbers of subjects and different ranges of measurements for each subject. For some combination of these two parameters, we counted the empirical probability of coverage of the real value $\boldsymbol{\alpha}$ from 10000 simulations taken by this confidence region. We considered model (1) with 2-dimensional regression parameter $\boldsymbol{\alpha} = (1, 2)'$, 2-dimensional vector of individual effects $\boldsymbol{b} = (b_1, b_2)'$ with covariance matrix $\mathbf{D} = (1, 0; 0, 1)$, AR(1) parameter $\rho = 0.5$ and errors variance $\sigma^2 = 1$. Results are shown in the Tables 1-2.

Table 1. Simulated probability of coverage of 95% confidence region evaluated from (9) for different numbers of subjects with the same range of the repeated measurements on each subject.

Number of subjects	Range of the repeated measurements on each subject	Probability of coverage
5	5	0.8297
10	5	0.8797
30	5	0.9394
50	5	0.9384
100	5	0.9413
500	5	0.9510
1000	5	0.9586

Table 2. Simulated probability of coverage of 95% confidence region evaluated from (9) for small number of subjects with a different range of the repeated measurements on each subject.

Number of subjects	Size of the repeated measurements on each subject	Probability of coverage
5	5	0.8297
5	30	0.8687
5	50	0.8805
5	80	0.8933
5	100	0.9040
5	150	0.9082

4. Discussion and conclusions

As it turns out, that our "naive" concept of the confidence region is particularly suitable for large numbers of subjects, which shows the Table 1. It is also appropriate to note, that for

small numbers of subjects (5 or 10) is this conference region unsuitable or (30 and 50) liberal. However, for a sufficient number of subjects (from 100 subjects) also for small numbers of the repeated measurements on each subject proposed confidence region is approaching the theoretical value. Moreover, from the Table 2 can be concluded that despite the increasing number of repeated measurements on each subject for a small sample of the subjects approaching simulated probability of coverage is very slow to the theoretical value. It is caused because the proposed confidence region does not take into account the uncertainty inherent in estimating the variance-covariance parameters. This can be removed using the F distribution instead of χ^2 distribution, but there arise practical problems with the numbers of degrees of freedom for this F distribution, where are used different approximation, see e.g. [4] and [6], [7]. Unfortunately, these confidence regions were not yet studied in detail for the analysis of longitudinal data. Therefore we think that it would be appropriate, on the basis of additional simulations, to verify their properties, or to propose their improvement.

Acknowledgements

The research was supported by the grants VEGA 1/0077/09, APVV-SK-AT-0003-09 and APVV-RPEU-0008-06.

References

[1] Chi EM, Reinsel GC. Model for Longitudinal Data with Random Effects and AR(1) Errors. *Journal of the American Statistical Association,* 84 (406): 452-459, 1989.

[2] Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of Longitudinal Data. Oxford University Press, 2002

[3] Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. John Wiley & Sons, New Jersey, 2004.

[4] Harville D. Accounting for Estimation of Variance and Covariance in Prediction Under general Linear Model: An Overview. *Tatra Mountains Mathematical Publication,* 39: 1-15

[5] Hedeker D, Gibbons RD. Longitudinal Data Analysis. John Wiley & Sons, New Jersey, 2006.

[6] Kenward MG, Roger JH. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics,* 53 (3): 983-997, 1997

[7] Kenward MG, Roger JH. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis,* doi:10.1016/j.csda.2008.12.013, 2009

[8] Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. *Biometrics,* 38 (4): 963-974, 1982.