Chapter 1

Extension of the LMG formula for longitudinal data

In the following chapter some extenstions of the LMG formula beyond the simple linear regression model are shown. The focus is on repeated measurements models. These models extend the simple linear regression by allowing intra-subject correlation between repeated measures.

The dependence of within-subject measurements can be modeled by including random effects (mixed model) or by assuming correlated errors within a subject (marginal model). A mixed model can be extended by including a random slope per subject, allowing for less restrictive longitudinal shapes. The marginal approach can get more freedom by different specified covariance matrices of the error terms. An unstructured covariance matrix, where no restriction are imposed, allows for the most freedom. However, depending on the number of repeated measuresments and the sample size the covariance matrix can get too large to make reasonable inference about it (Fitzmaurice *et al.*, 2011).

The extenstion of the LMG formula in the Bayesian framework to longitudinal models is restricted to models where the conditional variance formula can be applied to get the explained variance of the submodel from the regression parameters of the full model. The focus is therefore on the fixed predictors and not on the random effects. The conditional variance formula can be used in the marginal models, where only the fixed effects are modelled anyway. In the mixed model framework, the conditional variance formula is applicable to random intercepts models. For random-slope models there are at least some difficulties involved, if it is possible at all, to get the explained variance of the submodel. This chapter shows the Bayesian LMG Implementation on a random intercept model and on a repeated measurement model with an unstructured covariance matrix.

1.1 random intercept model

The first example concerns a simple random intercept model with time-varying predictors. Different R^2 metrics exist for linear mixed models. The variance of a random intercept model with regression parameter $\boldsymbol{\beta}$ can be written as

$$\mathrm{Var}(y) = \sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2, \tag{1.1}$$

where $\sigma_f^2 = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}$, σ_α^2 is the random intercept and σ_ϵ^2 is the error term (Nakagawa and Schielzeth, 2013).

An R^2 that is guaranteed to be positive is defined in Nakagawa and Schielzeth (2013) as

$$\mathrm{R}^2_{\mathrm{LMM}} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}. \tag{1.2}$$

It is theoretically possible that the $\mathrm{R}^2_{\mathrm{LMM}}$ decreases when adding predictors (Nakagawa and Schielzeth, 2013). By adding predictors σ_f^2 should always increase and σ_ϵ^2 decrease. However, the σ_α^2 may also increase a little bit and the total R^2 may then be a little bit lower. The R^2 can not decrease by using the conditional variance formula on the full model to calculate the R^2 of the submodels, because the total variance is fixed. The results should be the same, as if we would fit a new model by maximum likelihood for each submodel and compare the explained variance of the fixed effects to the explained variance of the full model. In the Bayesian framework, the conditional variance formula is needed to account for the interdepdence of the submodels to each other. The total variance of the full model can be calculated as $\mathrm{Var}(y) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}b) + \sigma^2$ or by using samples of σ_α^2 as in (1.1). The error term could again be sampled or calculated as in (**??**). In the following examples, (1.1) is used and σ_α^2 and σ_ϵ^2 are sampled from their posterior distribution.

In repeated measurement studies, the focus is often in within-subject changes. The between-subject variance, estimated with the random intercept term, is of minor importance. The more important question may be, how much variance the fixed predictors explain, compared to the within subject error, which is

$$\mathrm{R}^2_{\mathrm{repeated}} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\epsilon^2}, \tag{1.3}$$

The square root of this term is known under the name correlation within subjects in Bland and Altman (1995). Often, there are between-subject and within-subject predictors in a model. If we are interested in the within-subject effect, we can use a model including only the between-subject predictors as the null-model.

The following example shows a simple random intercept model with time-varying predictors. The main question is which within-subject predictors are the most important ones. The between-subject variance is of minor importance.

The data are simulated from the following regression setting with $m = 4$ timepoints and $n = 20$ number of subjects ,

$$\begin{aligned} Y_{i,j} \sim \mathcal{N}(\beta_0 + x_{1_{i,j}}\beta_1 + x_{2_{i,j}}\beta_2 + x_{3_{i,j}}\beta_3 + x_{4_{i,j}}\beta_4 + \alpha_i, \sigma^2), \qquad & i = 1, \ldots, n \\ & j = 1, \ldots, m \end{aligned}$$

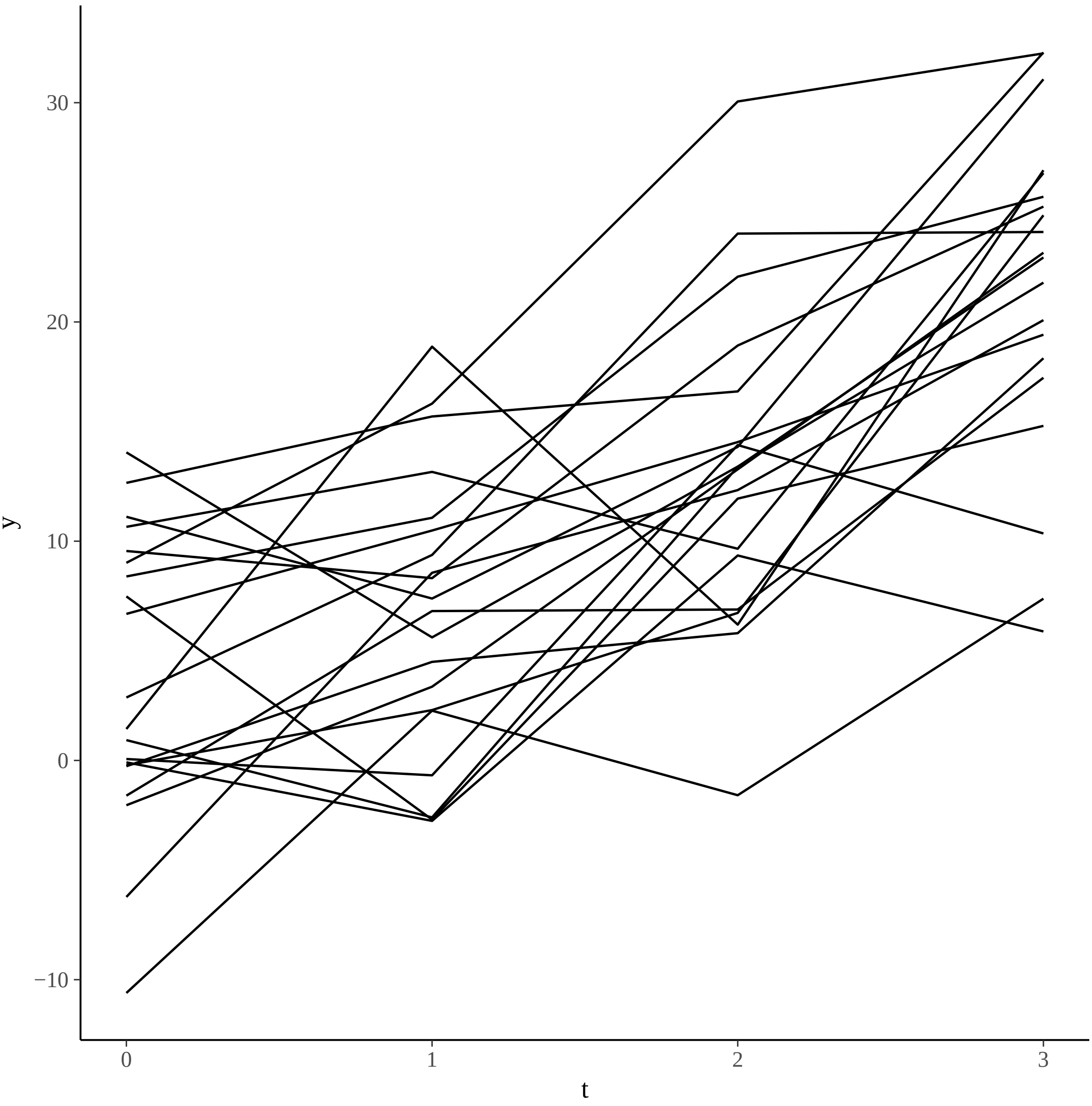

Figure 1.1: Individual trajectories of simulated random intercept model

where $\beta_1 = 1\,,\ \beta_2 = 1\,,\ \beta_3 = 2\ \ \beta_4 = 2\,,\ \sigma^2 = 1\,,\ \alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)\,,\ \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

The R code can be found in the appendix.

The random intercept effect is of minor interest. The R^2 of the models is calculated according to the formula of repeated measure correlation (1.3). Most of the within-subject variance is explained by the predictors (Table 1.1). The credible intervals are very narrow. For information about the between-subject variance term, we can look at the posterior distribution of the random intercept variance term.

In the second part, the random intercept is directly included in the total variance calculation of the R^2 values. There is a large between-subject variance in this simulated dataset (Table 1.2). The LMG values including the between subject variance are therefore much lower. The credible intervals are as well much wider, because the uncertainty about the between-subject variance is included.

In my opinion we can get more useful information from separating the between-subject and within-subject variance components in this simple case. Note that we assumed non stochastic predictors. Otherwise, the credible intervals would be larger. In general, it seems more reasonable to assume stochastic time-varying predictors. The variance could then be estimated by non-parameteric bootstrap, resampling whole subjects (all repeated measurements of a subject).

Table 1.1: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	I	J	Total
x1	0.202 (0.2, 0.203)	0.525 (0.523, 0.526)	0.727 (0.723, 0.73)
x2	0.21 (0.208, 0.211)	0.538 (0.536, 0.539)	0.747 (0.744, 0.751)
x3	0.297 (0.296, 0.299)	0.658 (0.658, 0.658)	0.955 (0.954, 0.957)
x4	0.291 (0.29, 0.292)	0.651 (0.65, 0.651)	0.942 (0.94, 0.943)

Table 1.2: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	I	J	Total
x1	0.108 (0.046, 0.172)	0.238 (0.116, 0.35)	0.346 (0.163, 0.522)
x2	0.089 (0.046, 0.127)	0.229 (0.116, 0.329)	0.318 (0.162, 0.456)
x3	0.113 (0.06, 0.161)	0.272 (0.139, 0.391)	0.385 (0.199, 0.551)
x4	0.115 (0.061, 0.163)	0.27 (0.137, 0.387)	0.385 (0.199, 0.55)

1.2 marginal model

The next example concerns a repeated measurement model with time-varying predictors and an unstructured error covariance matrix. The data are generated from the following model:

$$Y_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n \tag{1.4}$$

where $\boldsymbol{\Sigma}$ represents an unstructured error covariance matrix, $\mathbf{X}_i$ represents the predictor matrix of size $m \times p$ of subject i.

In the variance calculation we need to take into account that we do not have just one σ^2 parameter, but a covariance matrix $\boldsymbol{\Sigma}$. The diagonal elements of $\boldsymbol{\Sigma}$ represent the variance of each timepoint. The sum of the diagonal elements of $\boldsymbol{\Sigma}$ represents the variance for a whole subject. We can take the mean of diag($\boldsymbol{\Sigma}$) to make the formula compatiable with the $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}}\boldsymbol{\beta}$ of (**??**), resulting in the total variance term

$$\text{Var}(\mathbf{Y}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}}\boldsymbol{\beta} + \text{mean}(\text{diag}(\boldsymbol{\Sigma})), \tag{1.5}$$

The following R-code is used to generate the data:

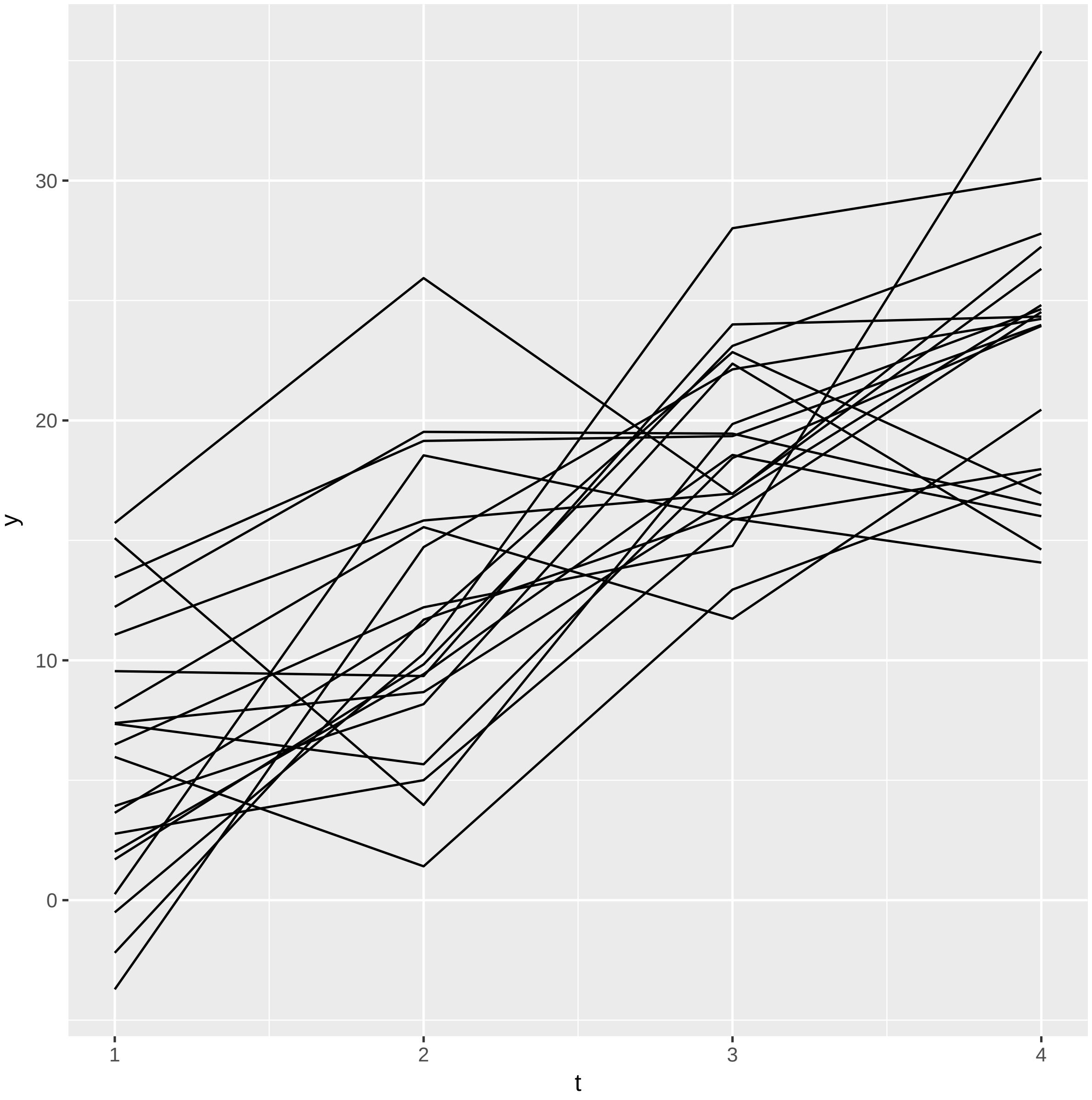

Figure 1.2: Individual trajectories of simulated data with unstructured error covariance matrix

The individual trajectories are shown in Figure 1.2. The resulting LMG values of the predictors are shown in Table 1.3.

Table 1.3: Variance decomposition for non-stochastic predictors. I = LMG values, J = joint contribution, Total = total explained variance in one-predictor only model

Variable	**I**	**J**	**Total**
x1	0.141 (0.116, 0.164)	0.365 (0.301, 0.41)	0.506 (0.415, 0.573)
x2	0.137 (0.11, 0.163)	0.355 (0.288, 0.4)	0.492 (0.4, 0.561)
x3	0.206 (0.169, 0.234)	0.445 (0.364, 0.495)	0.651 (0.533, 0.726)
x4	0.21 (0.171, 0.241)	0.445 (0.366, 0.496)	0.656 (0.539, 0.736)

Bibliography

Bland, J. M. and Altman, D. G. (1995). Calculating correlation coefficients with repeated observations: Part 1–Correlation within subjects. *BMJ (Clinical research ed.)*, **310**, 446. 2

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied longitudinal analysis*. Wiley. 1

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142. 2