Chapter 1

Extension to longitudinal data

Some extensions of the LMG formula beyond the simple linear regression model are shown in the following chapter. The focus is on repeated measures model. 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. Different covariance matrices of the error terms allow for less restrictions in the marginal approach. An unstructured covariance matrix, where no restriction are imposed, allows for the most freedom in the error term. However, depending on the number of repeated measurements and the sample size, the covariance matrix can get too large to make reasonable inference about it (Fitzmaurice *et al.*, 2011).

The extension of the LMG formula in the Bayesian framework applied 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. Therefore, the focus is on the fixed predictors, but not on the random effects. The conditional variance formula can be used in the marginal models, because only the fixed effects are modeled anyway. In the mixed model framework, the conditional variance formula is applicable to random intercept 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 submodels.

In this chapter, the Bayesian LMG Implementation is shown 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

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

1

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

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

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

Theoretically, it is possible that the R^2_{LMM} decreases when adding predictors (Nakagawa and Schielzeth, 2013). By adding predictors, σ_f^2 should always increase and σ_ϵ^2 should decrease. However, the σ_α^2 may also increase minimally and the total R^2 may then be minimally 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. Equal results should be obtained, by fitting a new model by maximum likelihood for each submodel and by comparing the explained variance of the fixed effects to the total variance of the full-model. In the Bayesian framework, the conditional variance formula is necessary to account for the mutual interdependence of the submodels. The total variance of the full-model can be calculated as $\text{Var}(y) = \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}b) + \sigma^2$ or by using samples of σ_α^2 as in definition (1.1). The error term could again be sampled or calculated as in definition (??). In the following examples, definition (1.1) was used, σ_α^2 and σ_ϵ^2 were sampled from their posterior distribution.

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

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

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 the interest lies in the within-subject effects, a model including only the between-subject predictors can be used 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 (see Appendix ?? for R-code),

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

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

The random intercept effect was of minor interest. The R^2 of the models was calculated according to the formula of repeated measure correlation (1.3). The R-code of the odel and the
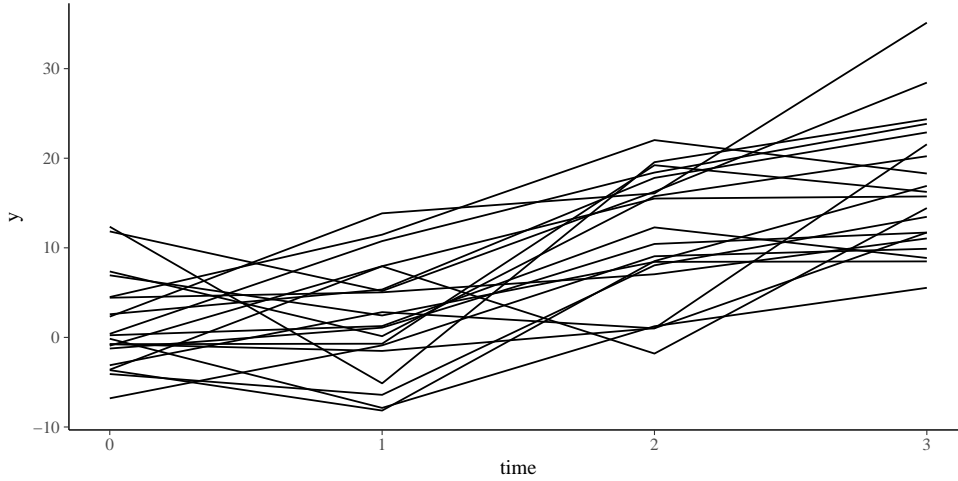
Figure 1.1: Individual trajectories of simulated random intercept model.

function can be found in the Appendix (**??** and **??**). Most of the within-subject variance was explained by the predictors (Table 1.1). The credible intervals were very small. For information about the between-subject variance term, we can look at the posterior distribution of the random intercept variance term.

Next, the random intercept was directly included in the total variance calculation of the R^2 values. There was a large between-subject variance in this simulated dataset (Table 1.2). Therefore, the LMG values including the between subject variance were much lower. The credible intervals were as well much larger, because the uncertainty about the between-subject variance was included.

In my opinion, we can get more useful information from separating the between-subject variance from the within-subject variance components in this simple case. Note, that we assumed non-stochastic predictors. Otherwise, the credible intervals were larger. In general, it seems more reasonable to assume stochastic time-varying predictors. The variance could then be estimated by non-parametric 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 model

Variable	I	J	Total
x1	0.189 (0.187, 0.19)	0.467 (0.465, 0.468)	0.655 (0.652, 0.659)
x2	0.19 (0.188, 0.191)	0.471 (0.469, 0.472)	0.661 (0.657, 0.664)
x3	0.311 (0.31, 0.312)	0.633 (0.632, 0.633)	0.944 (0.942, 0.945)
x4	0.31 (0.309, 0.311)	0.634 (0.634, 0.634)	0.944 (0.943, 0.946)

1.2 Marginal model

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

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

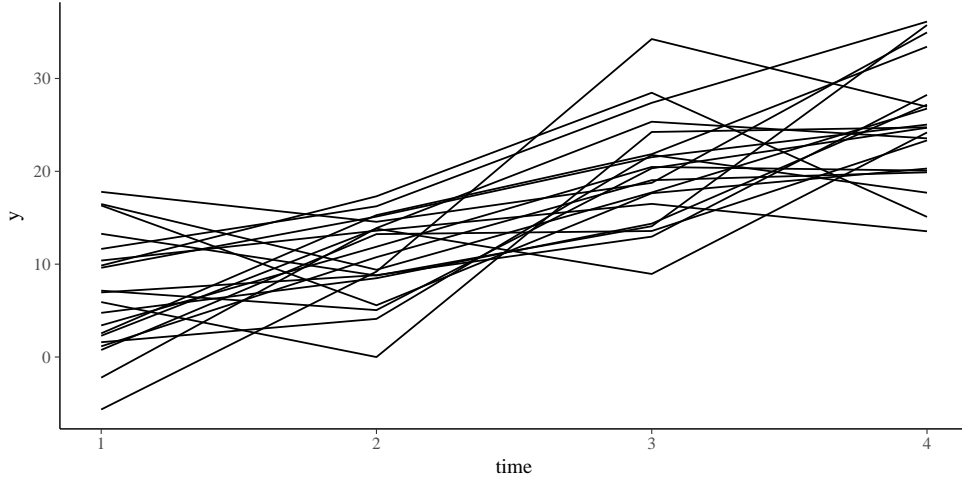| Variable	I	J	Total
x1	0.037 (0.019, 0.068)	0.079 (-0.011, 0.185)	0.112 (0.023, 0.251)
x2	0.066 (0.031, 0.108)	0.137 (0.055, 0.234)	0.204 (0.086, 0.341)
x3	0.104 (0.051, 0.165)	0.182 (0.075, 0.309)	0.287 (0.127, 0.473)
x4	0.114 (0.057, 0.178)	0.186 (0.077, 0.314)	0.301 (0.137, 0.491)

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

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

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

The R-code of the data generation can be found in Appendix **??**. In the variance calculation, it is necessary to take into account that there is not 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. The mean can be taken of diag($\boldsymbol{\Sigma}$) to make the formula compatible with the $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\beta}$ of (**??**), resulting in the total variance term

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

The individual trajectories are shown in Figure 1.2. The R-code for the model and LMG implementation can be found in Appendix (**??** and **??**). 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 model

Variable	I	J	Total
x1	0.154 (0.127, 0.174)	0.409 (0.337, 0.456)	0.563 (0.462, 0.63)
x2	0.155 (0.126, 0.178)	0.39 (0.322, 0.433)	0.544 (0.448, 0.609)
x3	0.213 (0.177, 0.237)	0.481 (0.401, 0.532)	0.695 (0.582, 0.768)
x4	0.206 (0.172, 0.228)	0.486 (0.404, 0.536)	0.692 (0.58, 0.764)

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