***Testing assumptions of mixed models: residual and influence diagnostics***

***Goal:*** 1) Introduce simulation-based graphical tools for diagnostics to assess if the underlying assumptions of mixed models are met. 2) Explore the problems associated with violations of the different types of assumptions.

**Step 1:** Assumptions about the residuals

**Sub-goal:** To get an overview about the assumptions that correspond to the distribution of residuals, provide graphical means to investigate how these assumption issues can be detected in the data and model at hand, and to demonstrate how the violation of these assumptions influence the estimated model parameters.

**Introduction**: Mixed models rely on several assumptions that should be fulfilled for deriving meaningful outputs from these models. Making inferences from models that violate these assumptions might be misleading. One set of assumptions corresponds to the residuals of the model that will be the focus of the current module. To understand the essence of residual diagnostics in the mixed model framework, we will first briefly describe the types of residuals one encounters when working with hierarchically structured data.

The term residuals commonly refer to the mismatch between the observed and predicted values by the model. For mixed models, we can define several types of residuals that should be considered for the verification of assumptions. Considering a random intercept model along the equation:

we incorporate uncertainty at two hierarchical levels, i.e., at the observation level () and at the individual level (). At the lower level, we can define level-1 residuals as the difference between what the model predicts and what we actually observe for individual *i* at instance *h* (this is lumped within ). At the higher hierarchical level, level-2 residuals define the deviations of each individual’s intercept from the population mean (as described by ). If the model also contains random slopes, these also are the components of the level-2 residuals. If there are higher hierarchical structures, we can further define higher-level residuals. Note that in essence, all higher-level residuals are the best linear unbiased predictors (BLUPs) of the random effects. In the mixed model framework, there are other types of residuals that are the linear combinations of level-1 and level-2 residuals, but these will be not considered here.

The basic assumptions in relation to the residuals can be summarized with the following simple equations:

These equations mean, respectively, that i) level-1 residuals are normally distributed around a mean of zero and are also homogenously distributed across groups of level-2 (individuals); ii) level-1 residuals are independent of each other; iii) level-2 residuals are also normally distributed around zero; iv) are independent of environmental covariates () that vary at the between-individual level; v) and independent of each other; and finally vi) there is no covariance between level-1 and level-2 residuals. Violations of these assumptions can be caused by biologically relevant processes. For example, if certain individuals are sampled in a specific range of the environment while others are sampled in another range (among individual variation in the environment) will create a scenario where there is a correlation between random intercepts (level-2 residuals) and the environment (x). Furthermore, it can happen that the random intercepts are not independent because individuals are genetically related. Violations of the assumptions may also be caused by missing an important variable affecting the phenotype of individuals. For instance, if we fail to control for sex in species where there is large sexual dimorphism in their behavior, the random intercepts for each individual may not follow a normal distribution.

Violations of the underlying assumptions of mixed models can cause problems in terms of the bias and precision of the estimated parameters. Several diagnostic tools can be used to determine if assumptions are violated allowing researchers not only to assess the quality of parameter estimation, but also to explore the features of the data that are not well described by the model. Therefore, evaluating whether the assumptions of the considered model are met and subsequently taking remedial actions, if necessary, may help researchers finding more appropriate models for their data. Residual analyzes are commonly used diagnostic tools for classical linear models including mixed effects. In this module, we will provide an overview of these tools.

**Exercise 1:**

For simplicity, we will continue relying on the basic mixed model that includes random intercept only:

The problems introduced here can easily expanded to models that include random slopes or contain more random effects on different hierarchical level.

**R code:**

# install.packages("lme4")

LMM1 <- lme4::lmer(Phenotype ~ 1 + X1 + (1|Individual), data = sampled\_data)

We will generate data according to this model in the same way as we did in the module for random intercept model, However, here we provide an additional option by which the user is allowed to vary the number of records per individual. To explore the effect of the number of samples per individual, we will try to establish scenarios in which there is a large difference in sample size among individuals! It is important to note that models apply to these data are not supposed to violate any of the assumptions.

Number of individuals: …….

Minimum number of trait expressions within-individuals (minimum 2): …….

Maximum number of trait expressions within-individuals (maximum 100): …….

Among-individual variance in the number of trait expressions within-individuals (0: all individuals have the same simple size, 1: maximum possible difference in sample size among individuals: …….

The rest of the parameter settings are the same as in the “random-intercept” module:

Among-individual variance (VI): …….

Variance due to population mean effect of an environmental factor ():…….

Residual variance (VR): …….

Note: you might remember that the residual variance can also include some unknown environmental variance in addition to the above-defined environmental variance; this will be important in the current exercise.

RUNNN

**Results**:

A mixed-effects statistical model estimates these parameters:

|  |  |
| --- | --- |
| True | Estimated |
| Fixed effects | Fixed effects |
| Mean of the trait (β0) = 0 | Sampled mean of the trait (β’0) = ….. |
| Population-specific slope of the environmental effect (β) = ….. | |  | | --- | | Estimated population-specific slope of the environmental effect (β’) = ….. | |
| Random effects | Random effects |
| Among-individual variance in intercept (VI) = …… | Sampled among-individual variance in intercept (V’I) = ….. |
| Residual variance (VR) = ….. | Residual variance of sample (V’R) = ….. |

For the purpose of diagnostics, we will examine some aspects of the distribution of residuals according to the assumptions of the underlying model. The upper panels are for the level-1 residuals (; histogram, individual-specific boxplots, residuals against fitted values and against predictor ). The lower panels show the level-2 residuals (; histogram, individual-specific intercept estimates with their errors). Note that one can also make similar plots on level-2 for the random slopes if they are also considered in the model (for simplicity, we did not include random slopes in this worked example). Furthermore, it is particularly useful to plot level-2 residuals against a predictor that varies at the among-individual level (e.g., , body size, sex).

GRAPHS in 2 (or 3) lines:

**1**, level-1 residuals: a) distribution (histogram or q-q plot or both); b) between-individual homogeneity (individual-specific boxplots); residuals against predicted; residuals against x

**2**, level-2 residuals: a) distribution (histogram or q-q plot or both); b) individual specific intercept (estimate and its error);

**Point:** Accidentally, these figures may show patterns that make you suspicions about the violated assumptions of the statistical model when fitted to the generated data, and you may think that some remedial action might be necessary. For example, if you have set a large variance for the within-individual sample size, the distribution of residuals across individuals may appear heterogeneous. However, you can be certain that the data were generated under the assumptions of the model, so any deviations from the null expectations of normality, homogeneity and independence in the above graphs should merely be the consequence of random noise in combination with the scarcity of data. To see the frequency by which such suspicious patterns can incidentally emerge in association with residuals that are coming from models and data that actually fulfil the underlying assumptions, click on the button below! By doing so, new data will be generated along the above settings (by using a fixed within-individual sample size (Hassen, please force the within-individual samples sizes to be the same as in the first run, so these are not generated again and again) and the same model will be fitted. The table of results and the residual plots will be updated. Try to repeat the data generation couple of times and investigate how the figures change. Remember, that everything you see here look normal – in a broad sense!

REFRESSSSSSSSSSS

**Exercise 2:** In the above example, you have seen how residuals should look when the underlying assumptions of mixed modelling were completely fulfilled. Now we will add an option, by which you can violate some of these assumptions and you can investigate the result of this intervention on the parameter estimates. You can also see how these effects are manifested in the plot of residuals. To be able to compare the outcomes, we will keep the parameters fixed as it were set in the above example. What you are allowed to vary is the type and degree of violation you impose on different assumption:

Level of investigation:

-level-1 residuals

-level-2 residuals

-correlation between level-1 and level-2 residuals

Type of violation:

-imposing non-normality

-imposing heterogeneity (uncontrolled predictor)

-imposing non-independence (autocorrelation)

Level of violation:

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

LOW INTERMEDIATE HIGH

RUNNN

**Results**:

A mixed-effects statistical model estimates these parameters:

|  |  |
| --- | --- |
| True | Estimated |
| Fixed effects | Fixed effects |
| Mean of the trait (β0) = 0 | Sampled mean of the trait (β’0) = ….. |
| Population-specific slope of the environmental effect (β) = ….. | |  | | --- | | Estimated population-specific slope of the environmental effect (β’) = ….. | |
| Random effects | Random effects |
| Among-individual variance in intercept (VI) = …… | Sampled among-individual variance in intercept (V’I) = ….. |
| Residual variance (VR) = ….. | Residual variance of sample (V’R) = ….. |

The distribution of residuals according to the assumptions of the underlying model. The upper panels are for the level-1 residuals (; histogram, individual-specific boxplots, residuals against fitted values and against predictor ). The lower panels show the level-2 residuals (; histogram, individual-specific intercept estimates with their errors).

GRAPHS in 2 (or 3) lines (same as above):

**1**, level-1 residuals: a) distribution (histogram or q-q plot or both); b) between-individual homogeneity (individual-specific boxplots); residuals against predicted; residuals against x

**2**, level-2 residuals: a) distribution (histogram or q-q plot or both); b) individual specific intercept (estimate and its error);

**Point:** Compare these outputs with the results of Exercise 2 (both the estimated parameters and the residuals plots). Can you see the difference between the residuals? Can you identify the violation that you imposed? Note the changes in parameter estimates.