# LMM or GLMM”

Random and fixed effects,

nested effects => level of analysis shouldn’t be pooled together ( don’t group all subjects 1 for example)

**Advantages**:

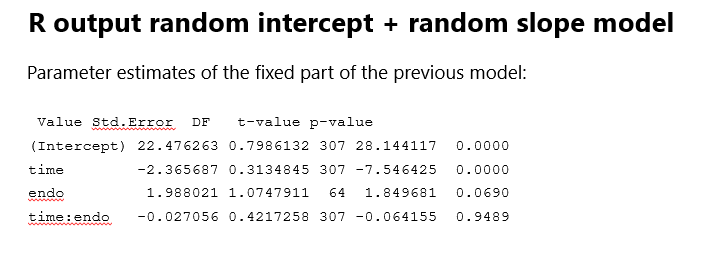
* Can handle missing data
* Handle unbalanced design
* Don’t need sphericity => variances around all levels need to be equal or close to equal

**Disadvantages**:

* Computationally more intensive
* Retain larger denominators of degree of freedoms => DF residuals

# Longitudinal data:

* Measures close together in time will be closer: week measures closer than month measures; We can check this using cor();
* Intercept represents all variables when the time = 0;
* Intercept represents difference in outcome when time = 0;
* Look table below:
  + ENDO x EXO:
  + As ENDO is defined the **reference is EXO**, so EXO starts with 22.47 (intercept)
  + Endo = intercept + slope => 22.47 + 1.98



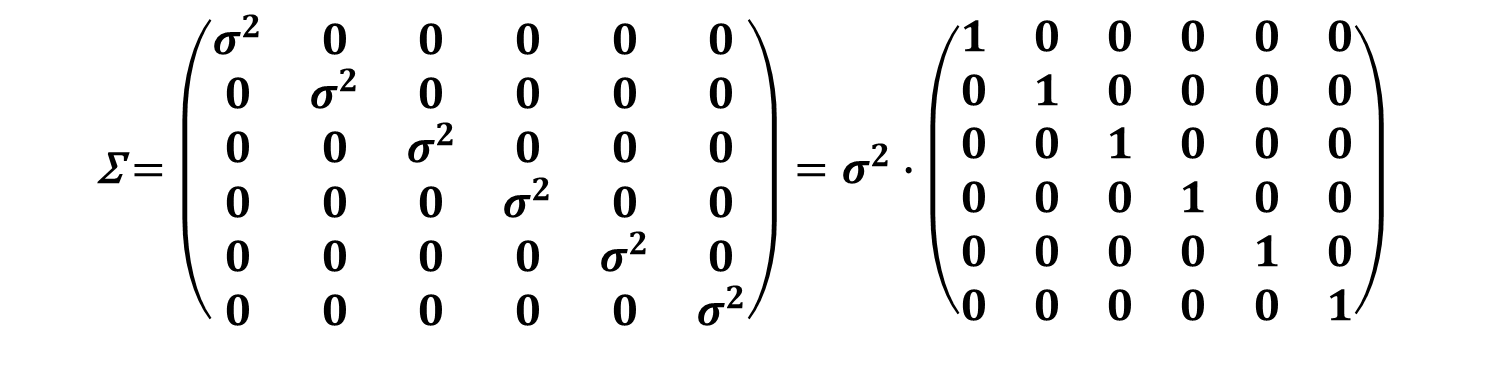
## LMM VAR – COR Matrix

* Model correlation of measurements *implicitly*
* Random intercept model implies a compound symmetry structure
* Random intercept and random slope also implies a certain correlation structure for the data => no simple structure
* structure depends on the estimates for , , and , but \*usually\* the variances increase for later time points and correlations decrease when time points are further apart

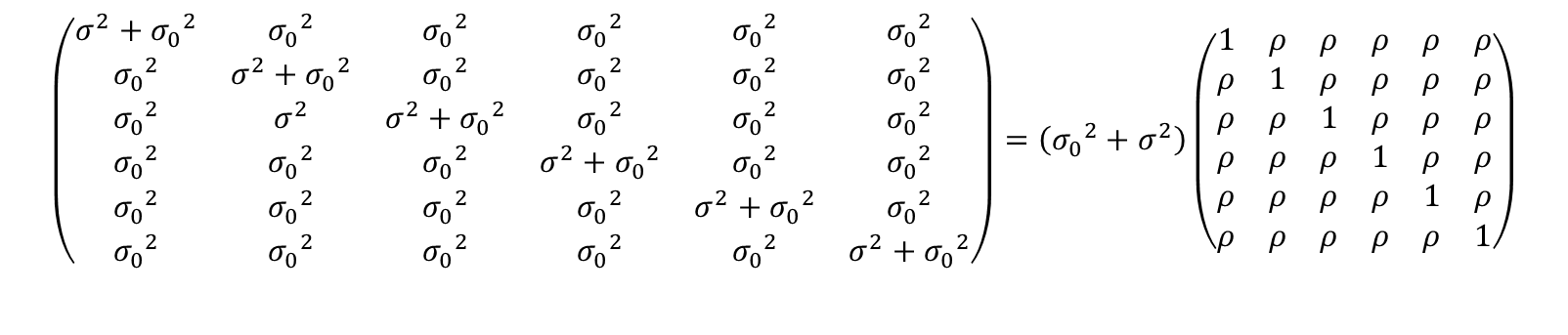
## CPM VAR-COR MATRIX (covariance pattern model or GEE-type cov structures)

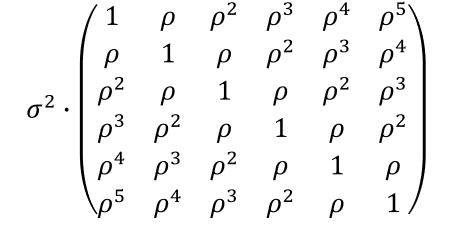
* No random effects
* Residuals are not independent => corr for ∑
* Var-cor (something complicated)
* Model correlation of measurements *explicitly*

1. Independent correlation structure
   1. The independent (scaled identity) correlation structure assumes residuals to be independent, as if they came from different subjects
   2. All variances are assumed equal, all correlations are assumed 0
   3. This is the assumption in ordinary linear regression/ANOVA



1. Compound symmetry correlation structure
   1. The compound symmetry (exchangeable) correlation structure assumes correlations between all time points to be equal, irrespective of the length of the time intervals.
   2. All variances are assumed equal, all correlations too:



1. Unstructured
   1. Var at each time point different
   2. Very expensive (cost a lot of degree of freedom)
2. Autoregressive of order 1: AR(1) (homogeneous)
   1. Assumes all observations 1 time unit apart have same correlations (p), 2 units corr (, and so on…
   2. Decreasing correlation over time.
   3. Outcome has same variance () across all time points
   4. 
3. Autoregressive of order 1: AR(1) (Heterogeneus)
   1. Allow var to differ over time
   2. Fits the data better

A random intercept model implies a compound symmetry structure for all data combined

A linear mixed model with random intercept and random slope also implies a certain correlation structure for the data, but this is by no means a simple structure

* + recall:
  + structure depends on the estimates for , , and , but \*usually\* the variances increase for later time points and correlations decrease when time points are further apart
  + this is exactly what we observed for our data set, so this model might fit the data quite well

# Summary

* Longitudinal data is a specific form of multilevel data
  + measurements within patients, challenge is in modelling time properly
* Time can be continuous or discrete
  + discrete: everyone measured at a few specific time points
    - but, with 3+ measurements per person and approximately linear time trends, you could still consider modelling data as continuous
  + continuous: measurements at different times for different individuals
* We can account for correlation of measurements over time
  + explicitly: variance-covariance matrix of residuals (CPMs)
    - primarily when everyone (theoretically) measured at same time points
  + implicitly: random intercept, random slope for time (LMEs)
  + (both explicitly & implicitly: LMEs with autocorrelated errors)
* “Baseline” measurement of outcome has different meaning depending on study design

# Testing in Linear Mixed Models

To decide which LMM fits the data best we can use likelihood- based methods:

* Likelihood Ratio Test (LRT) => LRT can be used to test nested models (one is a special case of the other) based on the χ²-distribution
* Akaikes Information Criterium (AIC) combination of likelihood and # parameters used in the model (d.f.) model with the lowest AIC (high likelihood with few parameters) is deemed best
* **Problem with ML estimation:**
  + variance parameters (residual variance, variance(s) of random effect(s)) **biased downwards (smaller than they really are!) => UNDERestimating the variance**
  + Divide by **n**(ML) or **n-1** (REML)
* **Solution: REstricted (or: REsidual) Maximum Likelihood (REML)**
  + gives unbiased estimates of variance parameters
  + BUT: adjusts likelihood for number of covariates in model, so cannot be used to compare models that differ w.r.t. fixed parts of model

# Technical Issues – Mixed Models

**When to use ML x REML:**

* Testing models that differ in variance components: REML will give interpretable LRT, AIC so will ML
* Testing models that differ in fixed effects: only ML will give interpretable LRT, AIC
* Leading me to suggest the following model-building strategy:
  1. Start with full fixed model and (using ML estimation), select appropriate random part of model
  2. With the random part chosen, (using ML estimation) try to reduce fixed part of model
  3. Once you have your final model: run that model once more using REML; this is the model you present to your audience
* Testing random effect(s):
  1. variance parameters are never <0
  2. LRT (REML/ML) for random effects: chi-square test, **but divide p-value by 2**
  3. AIC also okay
* Testing fixed effect(s):
  1. LRT (ML only!) for fixed effects: chi-square test, usual p-value
  2. AIC okay (only under ML)

**Checking assumptions of the model**

* Model assumptions:
  + linearity (if we use time – or other covariates – as linear)
    - check with individual plots, spaghetti plots, residual plots
  + normality of residuals
  + normality of random intercepts (& slopes, if used)
    - these three can be saved and checked using Q-Q plots, boxplots, histograms
    - but: generally not helpful
      1. because deviations from normality probably not a big problem for inference on fixed effects (if your interest is in inference on random effects, there could be a problem)
      2. model ‘inflicts’ normality on the random effects, so normality of the estimated random effects may partly reflect model assumptions
  + independence of residuals (once fixed and random effects are taken into account) **CANT CHECK**
    - as in linear models: keep your fingers crossed!

# **Generalized Linear Models**

* Data
  + Outcome variable Y
  + Predictor variable(s) X
* Model
  + Left-hand side: Y (continuous, dichotomous, count, ordinal, categorical, etc., from the exponential family)
  + Right-hand side: linear equation
  + Left- and right-hand side are linked together using an appropriate “link function”

## Example: logistic regression

* + Dichotomous outcome variable Y (1/0).
  + Link function: **logit**
  + Model:
* For example:
  + = pregnant (1 = yes, 0 = no), X = age, weight, LHB/CGB genes, etc.
  + = heart disease (1 = yes, 0 = no), X = age, weight, exercise, blood pressure, cholesterol
* **is the odds ratio corresponding to the effect of on**

## Example: Poisson regression

* + Count outcome variable Y.
  + Link function: **natural logarithm.**
  + Model:
  + For example:
  + Y = number of urinary tract infections per year, X = age, weight, antibiotics use, cranberry use, etc.
  + Y = number of telephone calls in NL on a given date, X = working day, season, temperature, economy, etc.
* Poisson regression: offset (extension of..)
  + Varying exposure window, e.g.
    - Insects (not all plots of land which we observe have the same size -> insects/km2).
    - Infections (not all patients were followed for the same length of time -> infections/year).
  + Formula:

* **is the rate ratio corresponding to the effect of on**

We can **interpret** the **Poisson regression** coefficient as follows: for a one unit change in the predictor variable, the difference in the logs of expected counts is expected to change by the respective **regression** coefficient, given the other predictor variables in the model are held constant.