# Epidemiology: week 2

<https://github.com/hansfranke1985/ADS/tree/master/DataWrangling/week_7_MissingData>

## Lecture 1: Types of missing data

## Ad-hoc Solutions

## 1. Listwise Deletion

* delete the missing values before the analysis (na.omit = True or complete.cases()
* may introduce additional complexities in interpretation
* Also known as complete case analysis (CCA)
* **Advantage**:

1. Convenience: If the data are MCAR, listwise deletion produces unbiased estimates of means, variances and regression weights.
2. Under MCAR, listwise deletion produces standard errors and significance levels that are correct for the reduced subset of data, but that are often larger relative to all available data.

* **Disadvantage**:

1. Wasteful
2. If the data are not MCAR, listwise deletion can severely bias estimates of means, regression coefficients and correlations.
3. Listwise deletion can introduce inconsistencies in reporting. Since listwise deletion is automatically applied to the active set of variables, different analyses on the same data are often based on different subsamples.
4. can lead to nonsensical subsamples ( p.e change the timeline when deleting days)

## 2. Pair-wise Deletion

also known as available-case analysis, attempts to remedy the data loss problem of listwise deletion. The method calculates the means and (co)variances on all observed data. Thus, the mean of variable X is based on all cases with observed data on X , the mean of variable Y uses all cases with observed Y -values, and so on.

* **Advantage**:

1. Under MCAR, it produces consistent estimates of mean, correlations and covariances

* **Disavantage**:

1. First, the estimates can be biased if the data are not MCAR. Further, the covariance and/or correlation matrix may not be positive definite, which is requirement for most multivariate procedures.
2. Problems are generally more severe for highly correlated variables
3. requires numerical data that follow an approximate normal distribution

## 3. Mean Imputation

Mean imputation should perhaps only be used as a rapid fix when a handful of values are missing, and it should be avoided in general.

* **Advantage**:

1. We may use for categorical data
2. Fast and simple

* **Disavantage**:

1. Mean imputation distorts the distribution in several ways
2. Understimate the variance, disturb the relation between variables, bias any estimation (other than mean)
3. Bias the estimation of the mean when not MCAR
4. Biases correlations to zero

## 4. Regression Imputation

Regression imputation incorporates knowledge of other variables with the idea of producing smarter imputations. The first step involves building a model from the observed data. Predictions for the incomplete cases are then calculated under the fitted model, and serve as replacements for the missing data.

* **Advantage**:

1. regression weights are unbiased under MAR if the factors that influence the missingness are part of the regression model

* **Disavantage**:

1. Imputing predicted values also has an effect on the correlation.
2. the ensemble of imputed values vary less than the observed values => variability of the imputed data is systematically underestimated
3. Note that this upward bias grows with the percent missing
4. In reality however, regression imputation artificially strengthens the relations in the data
5. is a recipe for false positive and spurious relations.

## 5. Stochastic regression imputation

Stochastic regression imputation is a refinement of regression imputation attempts to address correlation bias by adding noise to the predictions. This method first estimates the intercept, slope and residual variance under the linear model, then calculates the predicted value for each missing value, and adds a random draw from the residual to the prediction.

* **Advantage**:

1. A well-executed stochastic regression imputation preserves not only the regression weights, but also the correlation between variables
2. The main idea to draw from the residuals is very powerful, and forms the basis of more advanced imputation techniques

* **Disavantage**:

1. Can lead to strange results
2. Symmetric and constant error restrictive
3. Single imputation does not take uncertainty imputed data into account, and incorrectly treats them as real

## 6. LOCF and BOCF

Last observation carried forward (LOCF) and baseline observation carried forward (BOCF) are ad-hoc imputation methods for longitudinal data. The idea is to take the previous observed value as a replacement for the missing data. When multiple values are missing in succession, the method searches for the last observed value.

* **Advantage**:

1. LOCF is convenient because it generates a complete dataset

* **Disavantage**:

1. Implausive imputation (time series for example...)
2. LOCF can yield biased estimates even under MCAR

## Indicator Method

Suppose that we want to fit a regression, but there are missing values in one of the explanatory variables. The indicator method (Miettinen 1985, 232) replaces each missing value by a zero and extends the regression model by the response indicator. The procedure is applied to each incomplete variable. The user analyzes the extended model instead of the original.

* **Advantage**:

1. is that the indicator method retains the full dataset.
2. it allows for systematic differences between the observed and the unobserved data by inclusion of the response indicator, and could be more efficient.

* **Disavantage**:

1. The conditions under which the indicator method works may not be met in practice. For example, the method does not allow for missing data in the outcome, and generally fails in observational data.

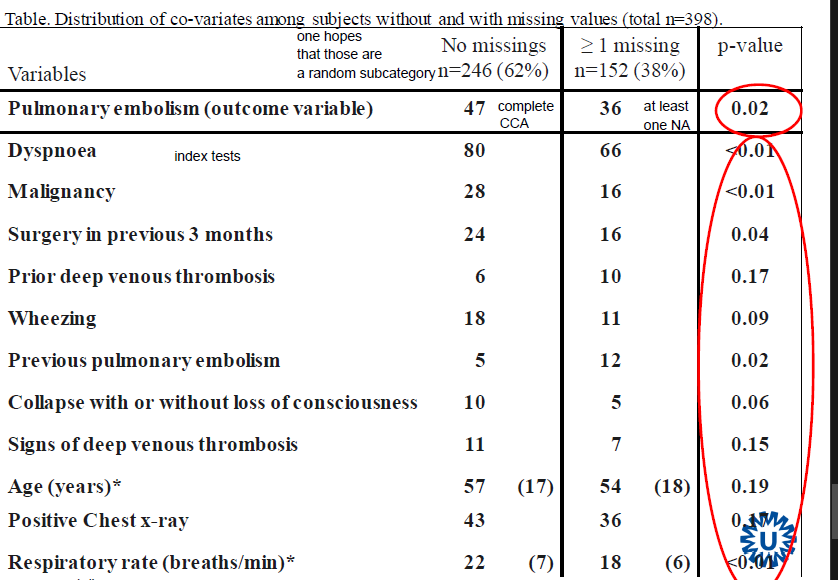
# Multiple Imputation

Multiple imputation creates m >1 complete datasets. Each of these datasets is analyzed by standard analysis software. The m results are pooled into a final point estimate plus standard error by pooling rules

1. Multiple imputation creates several complete versions of the data by replacing the missing values by plausible data values
2. estimate the parameters of interest from each imputed dataset. This is typically done by applying the analytic method that we would have used had the data been complete.
3. The last step is to pool the m parameter estimates into one estimate, and to estimate its variance. The variance combines the conventional sampling variance (within-imputation variance) and the extra variance caused by the missing data extra variance caused by the missing data (between-imputation variance).

Under the appropriate conditions, the pooled estimates are unbiased and have the correct statistical properties.

Another reason to use multiple imputation is that it separates the solution of the missing data problem from the solution of the complete-data problem. The missing-data problem is solved first, the complete-data problem next. Though these phases are not completely independent, the answer to the scientifically interesting question is not obscured anymore by the missing data. The ability to separate the two phases simplifies statistical modeling, and hence contributes to a better insight into the phenomenon of scientific study.



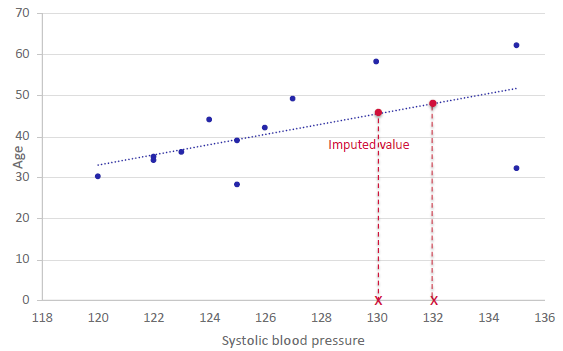
* if 246 would have exactly same values as the 152 it would be MAR and we could use CCA
* Values differ between group with missing values and group without missing values
* we have to assume that the difference is dependent on the missingness
* variables are associated with missingness

# Epidemiology: week 3

## Lecture 1: Imputation of Missing Data

**Single (multivariable) regression based imputation**

* Includes outcome! -> preserves relation
* Includes all variables of the model
* Includes unknown predictors of the missing value
* Estimate actual value of a missing value given all other predictors

****

**Benefit:**

Uses all information (covariates and outcome)

**Problem:**

imputed values directly on regression line compared to observed values being scattered

**-> lacking variation**

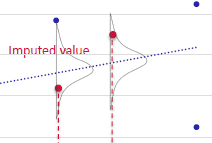
**Uncertainty of natural variation** **& uncertainty of estimated imputation model** (SE too low)

**Multivariate missing data** -> model cannot make use of observed data from other columns to calculate predicted values

**Solve problem by**:

**Adding natural variation** (binary case)

* Imputation is given by random sample from Bernoulli

**Adding natural variation** (continuous case)

* Adopt linear regression
* Yields residual variance
* Imputation by random sample from Normal
* adding variability by adding scatter based on the residual error -> imputed values NOT on regression line

**Standard Error is still underestimated (too significant)**

* we are **treating all data as if it was observed** data BUT the observed values are obviously more reliable than the imputed values and we are making an assumption by treating them the same

## Lecture 2: Imputation of Missing Data

**Multiple imputation by regression**

**MI by regression**

**recommendations:**

* should reflect all uncertainty (imputation model error & uncertainty on imputation model parameters)
* should be as flexible as possible (complexity for analysis model & interaction of model)
  + other variables may also carry information on NAs
* ALWAYS include the outcome (directly or indirectly)

**common pitfalls**:

* keeping output variable/ outcome out of the imputation
  + final analysis model is a model of the outcome predicted by several factors
  + **association between imputed predictor and outcome is lost when outcome is not included during the imputation**
    - leads to congeniality problem= the fact that all relations in analysis model should be represented in the imputation models
* non-normally distributed assumption (check setting of package)
* false assumption of MAR (e.g. falsely assuming that we have information to impute NAs)
* computational problems

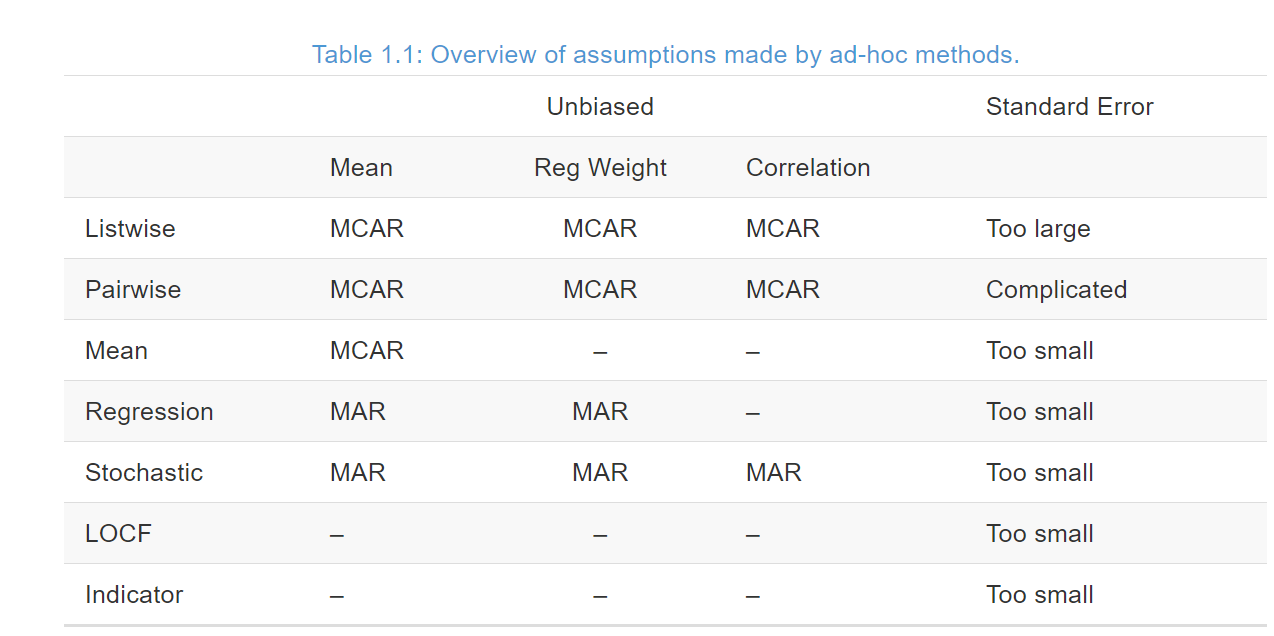
**Special case when only outcome data is missing & analysis based on max. likelihood:**

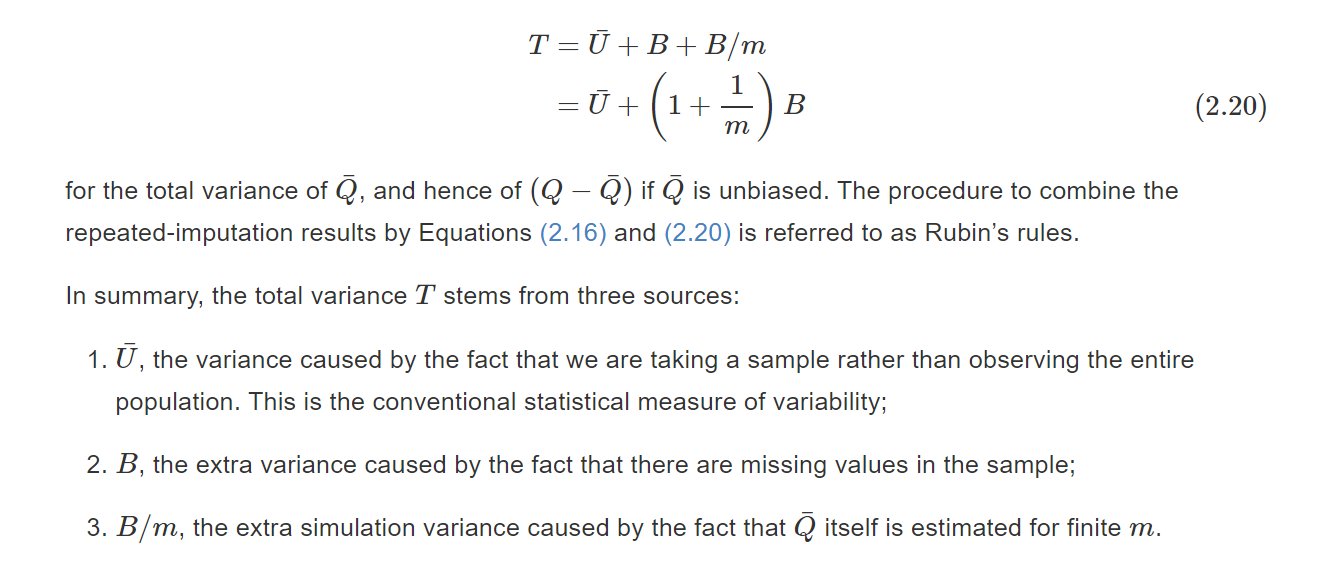
* unbiased & no imputation needed if:
  + CCA with adjusted covariate
  + Missing outcomes are MAR (NAs are conditional on data we have collected)
  + All covariates of outcome are included as covariates in the adjusted model

= fully adjusted model

**BUT**

* One can still reduce uncertainty by using imputation
* It allows for integration of post-randomization variable





**Q & A**

**Goals of imputation of missing data:**

* Maintain relationship in the data (correlation structure)
* Retain variation in the data
* Keep uncertainty in outcome that’s equal to the real uncertainty
* Keep Uncertainty in imputation model that’s equal to real uncertainty
  + Values of the coefficients for each covariance have to be reflected by imputation model
    - Draw values for each coefficient and generate imputations according to these values (beta)
* being able to impute missing values when NAs on multiple variables
  + input random value/mean and then iterate to improve the imputed values (there will always be uncertainty around the coefficient for each imputation)

**Goal is not to find best imputed values but to find the values that can be plotted best that maintain the variation and uncertainty of the given data**

**When should I use single imputation?**

Never 😊 The standard error is only smaller in single imputation because it is BIASED.

# Epidemiology: Mixed Models (week 4 and 5)

## How to analyze relation between exam score and LRT score?

1. **linear regression, mean exam per school vs mean LRT (“aggregated data”)**

**Disadvantages:**

* 1. every school (regardless of sample size) given equal weight
  2. N = 65
  3. school-level variables possible, but not child-level variables
  4. we can only make inference at school level, not child-level
  5. possibility of “ecological fallacy”

1. **linear regression, all schools together (“disaggregated data”)**

**Disadvantages**:

* 1. inflates sample size, especially for level-2 variables
  2. SE’s of level-2 variables tend to be underestimated → p-values too small, CI’s too narrow (type I error inflated)
  3. SE’s of level-1 variables may be over- or underestimated
  4. ignore correlated residuals (correlation of children within schools)

1. **linear regression per school**

**Disadvantages**:

* 1. 65 different regressions, how to combine the results?
  2. mean slope: every school has equal weight
  3. standard error of parameter estimate correct?
  4. child-level variables possible, but not school-level variables

1. **linear regression, all schools together, regression with main effect and interactions to allow for different intercepts and slopes**

**Advantage over previous analysis:**

* 1. now we can include both child- and school-level variables
  2. residuals probably normally distributed (with constant variance?) around individual lines

**Disadvantages:**

* 1. We wanted 1 intercept and 1 slope for LRT, but:
  2. 65 schools, so 1 reference category and 64 estimates for intercepts (main effects per school) + 64 estimates for interactions (slopes per school)!
     1. Which school is the reference?
  3. We can’t generalize beyond these 65 schools

1. **Linear Mixed Model**

**Advantages:**

* 1. sample size correct, account for correlation of children within schools
  2. so: correct SE’s/p-values/CI’s
  3. no need for 64 main effects and interactions
  4. differences between schools captured one or more ‘variance components’
  5. both child-level and school-level variables simultaneously
  6. so: inference for both children and schools
  7. interactions between child- and school-level variables possible
  8. examine variation at different levels
  9. models work well in presence of missing outcomes (longitudinal)

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**:

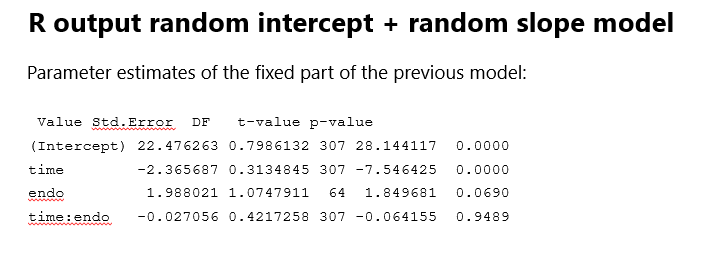
* Computationally more intensive
* Retain larger denominators of degree of freedoms => DF residuals

**Model assumptions:**

* + ; ;
  + independent

# Longitudinal data:

* variable measured on individual at several time points
* observations on one and the same individual will not be independent
* calls for special analysis techniques
* 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

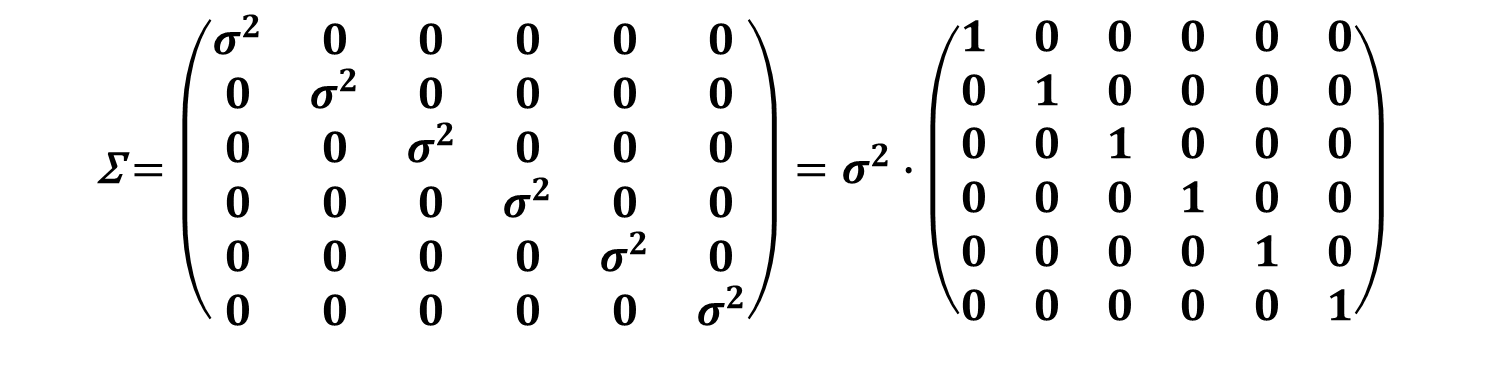
* Model correlation of measurements ***implicitly***
* Random intercept model implies a compound symmetry structure
* Random intercept and random slope also implies a certain correlation structure for the data => no simple structure
* structure depends on the estimates for , , and , but \*usually\* the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

## CPM – Covariance Pattern Model

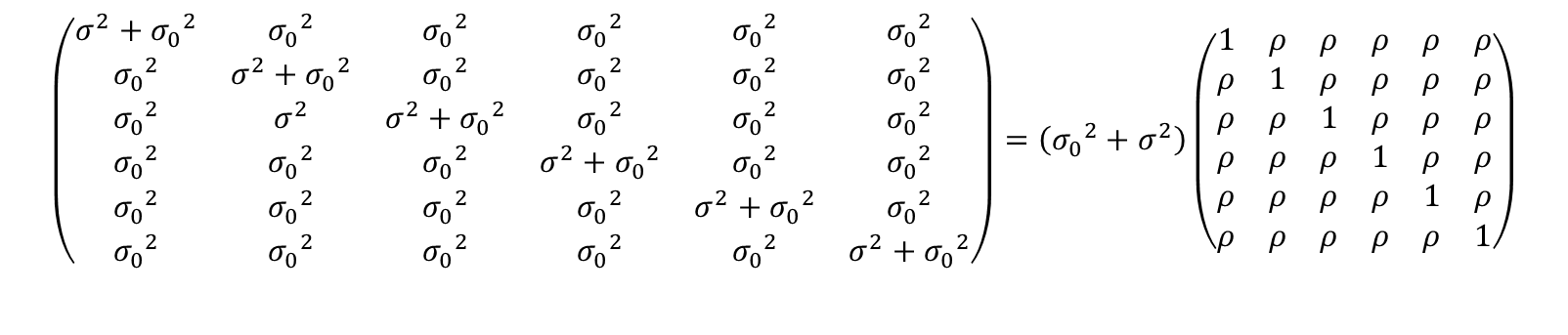
## CPM VAR-COR MATRIX (covariance pattern model or GEE-type cov structures)

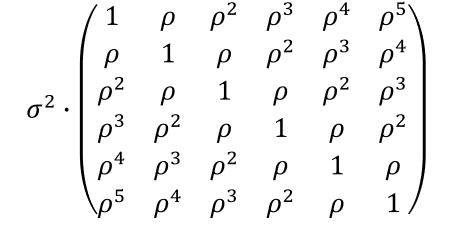
* No random effects
* Residuals are not independent => corr for ∑
* Var-cor (something complicated)
* Model correlation of measurements ***explicitly***

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:
   3. …is equivalent to a linear mixed model with a random intercept per patient



1. **Unstructured**
   1. Variances at each time point different / All covariances among time points different (note that θ12 = θ21)
   2. Very expensive (cost a lot of degree of freedom)
   3. Flexible structure
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 variance to differ over time
   2. Fits the data better

**“Covariance patterns” of linear mixed models:**

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!

# **- deviance is** -2 \* logLik, **- AIC is** -2\*logLik + 2\*npar

**What is the effect of centering an explanatory variable?**

* + changes the interpretation of the fixed intercept
  + can change the variance of the random intercepts, and the correlation of random intercepts with random slopes

# **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:

LOG RATE => RATE RATIO

* **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.