Applied Multivariate and Longitudinal Data Analysis

Longitudinal Data Analysis: Linear Mixed Model

Ana-Maria Staicu
SAS Hall 5220; 919-515-0644; astaicu@ncsu.edu

1 Introduction

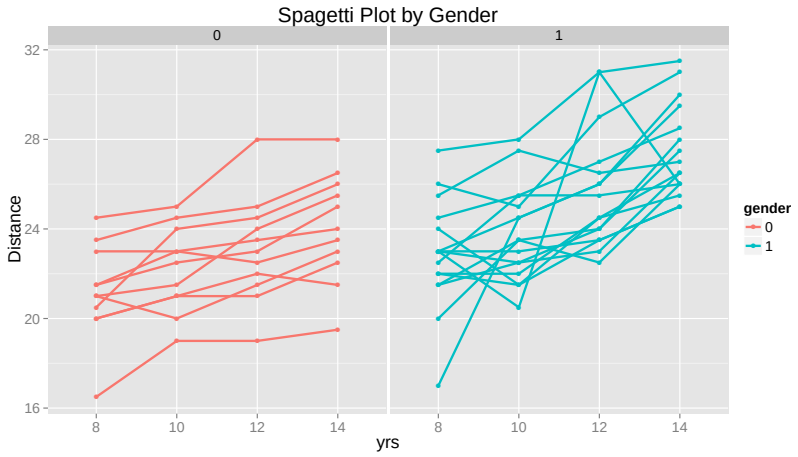
In the previous chapter we took a population perspective and modeled the response in terms of a population behavior and a random deviation from this. In contrast the modeling we study in this chapter focuses primarily on modeling the subject/unit trajectory. Hence the name "subject-specific" approach that is commonly used for this approach.

Intuition: consider the subject/unit trajectory itself and model its behavior using *2 stages*.

> 1st stage: Describe the trend of each subject trajectory by using some sort of parametric model and subject-specific parameters (*subject-level stage*)
>
> 2nd stage: Describe how the subject-specific parameters vary across subjects (*population stage*).

This modeling approach explicitly acknowledges the two sources of variation: within-unit and between-unit. This perspective offers more flexible models that do not require balanced/regular designs across units/subjects, allows for more general covariance structures, and can easily accommodate additional covariate information. In this chapter we will discuss interpretation of the model components/estimation and inference as well as prediction of the full subject trajectory.

Dental study. 'Dental growth measurements of the distance (mm) from the center of the pituitary gland to the pterygomaxillary fissure were obtained on 11 girls and 16 boys at 8, 10, 12, and 14.'



Random Coefficient Model (RCM)

Consider the dental study example and discuss the following:

- Examine the response trajectory for each subject in part. What features do you observe (look at how the distance varies over time and how it changes) ?

- Discuss a model that could be used to describe mathematically the way that the distance response for a subject varies over time. Make sure that whatever parameters you use are specific to the subject.

- Using this perspective, what type of variation do you think you can quantify at this step. Rrecall, we talked about two types of variations: within and between.

- What is the other type of variation that we need to describe. Intuitively try to suggest one way to describe this other type of variation.

Recall set up: $Y_{ij} = Y_i(t_{ij})$ denotes the response obserbed for the ith subject at time t_{ij}. A **linear mean trend** model for the response trajectory of the ith subject is specified by

$$Y_i(t_{ij}) = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}$$

where β_{0i} is the *subject-specific intercept* and β_{1i} is the *subject-specific slope*. The line $\beta_{0i} + \beta_{1i} t_{ij}$ describes the response trajectory for the ith subject; it is called the *subject-specific mean trajectory*. The departure of the response Y_{ij} from the ith subject-mean trajectory is considered random and is attributable to either biological fluctuations about the subject-mean trajectory or measurement error (or both). Here e_{ij} denotes the random deviation from the subject-mean trajectory; it is assumed to have zero-mean. The model assumed for the variation of e_{ij} describes the *within-unit* variation.

The parameters included in the specification of the subject-mean trajectory depend on the subject and thus are assumed random; let $\beta_i = (\beta_{0i}, \beta_{1i})^T$ denote the full parameter for the subject i. The model assumed for the variation of β_i describes the *between-unit* variation.

Because the subject level parameters β_i are random, this modeling approach is called 'random coefficient model' in short RCM. RCM assumes a parametric model for the subject mean trajectory; the linear model assumed above is just an example of RCM. Another example would be a **quadratic mean model** $Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij}$. RCM are a particular case of the wider class of models 'linear mixed effects (LME) models' which we study later in this chapter.

In class: Consider a linear mean model specified as above $Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}$, where i indexes the subjects and j indexes the repeated measures. Assume that the subject-level parameters β_i are normally distributed:

$$\begin{bmatrix} \beta_{0i} \\ \beta_{1i} \end{bmatrix} \sim iid\ N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \right).$$

- Re-write the model for Y_{ij} in a way that specifically describes the systematic trend in the population or population mean trend and the random deviation.

- Write down the population mean trajectory. Discuss whether this makes sense (recall the model used for the subject mean trajectory)

- Focus on the between-units random deviation. First discuss the variability measured by D. In particular discus 1) $D_{12} = 0$ and 2) $D_{11} = D_{22}$ What are the practical implications of these cases?

- Focus on the total random deviation. i) Let $e_i = (e_{i1}, \ldots, e_{in_i})^T$ be the vector of subject-level random deviations. Denote by $R_i = cov(e_i)$. Write down the total variation (within-unit + between-unit)

- Focus on the total random deviation. ii) To gain insight into the flexibility of this approach, consider next several particular cases. In each case write the model variance $\Sigma_i = cov(Y_i)$:

  a) Assume $R_i = \sigma^2 I_{m_i}$.

  b) Assume $R_i = \sigma_1^2 \Gamma_i$ where Γ_i is compound symmetric specified by parameter ρ.

  c) Assume $R_i = \sigma^2 I_{m_i} + \sigma_1^2 \Gamma_i$ where Γ_i is compound symmetric specified by ρ.

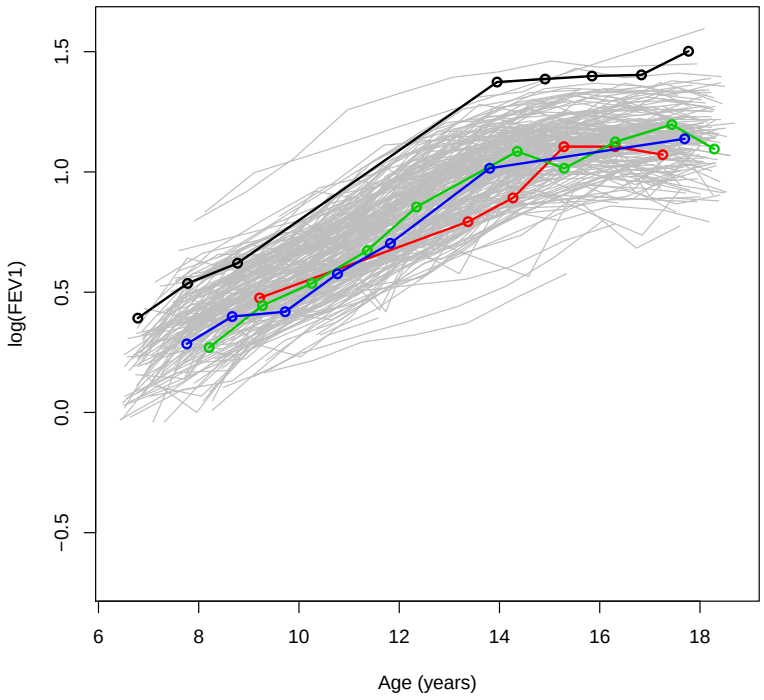
  What are the model parameters in each of these cases (mean parameters as well as covariance parameters)? Discuss your observations.

In general R_i is specified as $R_i = \sigma_1^2\Gamma_i + \sigma_2^2 I_{m_i}$, where Γ_i has some known correlation pattern model (AR1, exchangeable, Toeplitz, etc as described in the previous chaper). The two components describe the two sources of variability at the subject level: biological variation about the subject mean trend (which is quantified by $\sigma_1^2\Gamma_i$) and the measurement error (which is quantified by $\sigma_2^2 I_{m_i}$. In practice the structure assumed for R_i is related to which of the two sources is believed to dominate. Specifically an assumption like $R_i = \sigma^2 I_{m_i}$ has the interpretation that the measurement error dominates, while an assumption like $R_i = \sigma_1^2\Gamma_i$ means that the biological variation dominates. Taking $R_i = \sigma^2 I_{m_i}$ is very common and tends to be the default case for many statistical procedures.

In class example: The Six Cities Study of Air Pollution and Health

The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents, and the factors that influence lung function growth. A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the U.S.: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas). Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up. At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian.

The dataset contains a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of forced expiratory volume (FEV1), height (Ht) and age (Age) obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The random sample consists of 300 girls, with a minimum of one and a maximum of twelve observations over time.



Based on the plot above, write down a possible RCM to describe the variation of the FEV1 over time (Age) and accounts for the height (Ht) of the females.

Linear mixed effect model (LMM)

Setup: Observed data are repeated observations from n units/subjects/objects, indexed by i; $\{Y_{ij}, t_{ij},$ and other covariates : $j = 1, \ldots, m_i\}$. The *linear mixed effect model (LMM)* is the model obtained by aggregating the two levels below:

- (1st level) $Y_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i + e_{ij}$ with e_{ij} zero-mean random deviations. In vector format the LMM can be written:
$$Y_i = X_i\beta + Z_i b_i + e_i$$
  - β is K-dimensional vector of fixed parameters; X_i- $m_i \times K$ dimensional fixed design matrix with rows X_{ij}^T
  - b_i is q-dimensional vector of random parameters; Z_i- $m_i \times q$ dimensional random design matrix with rows Z_{ij}^T;
  - $e_i = (e_{i1}, \ldots, e_{im_i})^T$ is m_i-dimensional vector of residuals. It is assumed independent over i, $e_i \sim N_{m_i}(0, R_i)$.
- (2nd level) Describe the model for the subject-parameters
$$b_i \sim N_q(0, D)$$

Interpretation of the model parameters. Answer the following questions.

- (Conditional perspective) Write down the distribution of $Y_i|b_i$ (that is calculate the conditional mean and the conditional variance).
  - $E[Y_{ij}|b_i]$. The conditional mean describes the mean response of an individual/unit
  - $Var(Y_i|b_i)$
- (Marginal perspective) Write down the distribution of Y_i (that is calculate the marginal mean and the marginal variance).

– $E[Y_{ij}]$. The marginal mean describes the mean response in the population. For this reason the fixed parameter β is also referred to by "population-average " parameter. It has population average interpretation.

– $Var(Y_i)$

In class: Assume the model $Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i1} t_{ij} + e_{ij}$. Interpret the slope parameters β_1 and b_i.

Recall the general LMM model $Y_i = X_i\beta + Z_i b_i + e_i$ for $e_i \sim N_{m_i}(0, R_i)$ and $b_i \sim N_q(0, D)$. Furthermore assume that R_i has a known structure and depends on a vector parameter. **Model parameters:**

- Fixed:
  - mean regression parameters β
  - covariance regression parameters $\omega =$ vector of the elements of D and the parameters of R_i.
- Random:
  -subject specific random effects b_i.

In the following we discuss the estimation of the fixed effects and the *prediction* of the random effects.

Estimation/inference of model parameters

Intuition: Estimation and inference about the fixed effects is based on the marginal model for Y_i, and hence it follows the same general ideas as the estimation/inference in the population average models. Prediction of the fixed effects is based on the conditional model $Y_i|b_i$.

Fixed effects parameters

Recall the model

$$Y_i = X_i\beta + Z_i b_i + e_i,$$

where $b_i \sim N_q(0, D)$ and $e_i \sim N_{m_i}(0, R_i)$ with b_i independent of e_i. Denote by ω the vector of the covariance parameters (the parameters describing the covariances D and R_i).

Here we discuss statistical inference about the population-level parameter, β. Inference for the classical LMM focuses on the fixed effects parameters β. For these models estimation uses the methods based on the marginal log-likelihood. ML/REML are used to estimate both β and ω. Consequently the inferential tools are very similar to those introduced for the marginal models for correlated data.

The same methods for carrying inference for β apply:

- GLS to estimate β, denoted by $\widehat{\beta}$
- $var[\widehat{\beta}] = V_{\widehat{\beta}}(\omega)$ as before
- Distribution of $\widehat{\beta} \sim N_k(\beta, V_{\widehat{\beta}}(\omega))$. For large samples substitute the true covariance parameter ω by its estimate $\widehat{\omega}$. For notation we use $\widehat{V_{\widehat{\beta}}} = V_{\widehat{\beta}}(\widehat{\omega})$
- Testing hypotheses involving β is done by formulating the hypothesis as $L\beta = h$ and using Wald / LRT procedures studied earlier which we briefly review. If the null hypothesis is

  $$L\beta = 0$$

  then one can use Wald (with its $t_{\widehat{\nu}}$ counterpart for smaller sample size, and row vector L) or LRT. Wald and LRT have an asymptotically χ^2 distribution under the null hypothesis.

Random effects parameters

One main advantage of this methodology is the ability to predict individual trajectories, by incorporating individual specific information. The idea is to exploit the fact that the individual mean is described by $X_i\beta + Z_i b_i$. In particular if $\widehat{\beta}$ and $\widehat{b}_i$ are the estimates of β and b_i, then the ith individual trajectory can be 'predicted' by

$$\widehat{E[Y_i|b_i]} = X_i\widehat{\beta} + Z_i\widehat{b}_i$$

Terminology: *Predictor* is an estimator of a random quantity. In contrast, *estimator* is a term correctly used only in regards to a fixed unknown parameter.

Intuition: the subject specific random deviations b_i will be predicted using best linear unbiased prediction (BLUP). The predictors are:

i) linear functions of the data (hence the name *linear*)

ii) *unbiased* - that is the average value of the predictor is equal to the average (random) quantity being predicted;

iii) *best* - in the sense that they achieve minimum mean squared error among all the linear + unbiased predictors.

BLUPs were introduced by Henderson (1950) - as the 'joint maximum likelihood estimates'; although the actual name/acronym was used later by Goldberger (1962). He latter corrected this name on the basis that the function being maximized is not a proper likelihood function.

Approach 1. Assume that both b_i's and e_i's have normal distribution as above. The specific function we refer to is the joint density of b_i's and Y_i's:

$$\begin{aligned}\prod_{i=1}^{n}\{f(Y_i|b_i)\times f(b_i)\} &= \prod_{i=1}^{n}(2\pi)^{-m_1/2}|R_i|^{-1/2}\exp\{-\frac{1}{2}(Y_i-X_i\beta-Z_ib_i)^TR_i^{-1}(Y_i-X_i\beta-Z_ib_i)\}\\ &\times\prod_{i=1}^{n}(2\pi)^{-q/2}|D|^{-1/2}\exp(-\frac{1}{2}b_i^TD^{-1}b_i).\end{aligned}$$

The maximization is carried in the typical way, that is, by working in the logarithm scale, taking the (partial) derivatives with respect to β and b_i and setting them to zero.

We derive the solution $\tilde{\beta}$ and $\tilde{b}$ by using the form of the LMM $Y=X\beta+Zb+e$, where Y is the N-dimensional vector obtained by stacking all the responses Y_{ij} first over j and then over i, X is $N\times K$ full fixed design matrix, b is the nq dimensional vector obtained by column stacking the random deviations b_i, Z is the $N\times nq$ full random design matrix obtained by stacking the subject random design matrices Z_i in a diagonal pattern so that Zb is the column stacking of Z_ib_i over i's. Also e is the full vector of residuals. We denote by $R=\mathrm{diag}\{R_1,\ldots,R_n\}$ the block diagonal matrix of dimensions $N\times N$ with block matrices $R_1,\ldots R_n$. Also let $G=\{D,\ldots,D\}$ the block diagonal matrix of dimensions $nq\times nq$ with n diagonal elements $D=\mathrm{cov}(b_i)$.

Using this notation, the -2 $\times$ log of the joint density (to be minimized) can be written as:

$$(Y-X\beta-Zb)^TR^{-1}(Y-X\beta-Zb)+b^TG^{-1}b \tag{1}$$

by ignoring the constant terms with respect to β and b. Let's make the following notations:

$C=[X|Z]$ $\quad N\times(K+q)$ dimensional matrix obtained by concatenating the two design matrices

$B=\begin{bmatrix}0_{K\times K} & 0_{K\times nQ}\\ 0_{nQ\times K} & G^{-1}\end{bmatrix}$.

The solution can be written as:

$$\begin{bmatrix}\widetilde{\beta}\\ \widetilde{b}\end{bmatrix}=(C^TR^{-1}C+B)^{-1}C^TR^{-1}Y; \tag{2}$$

this solution is often referred to as *mixed model solution*. In practice this solution is not readily accessible as the model variance-covariances are not known and require estimation.

Approach 2. Recall that b is vector of random effects - what is a good prediction?

Let's consider a simple example;

$$y = v + \epsilon, \qquad \begin{bmatrix} v \\ \epsilon \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \right). \tag{3}$$

Assume we observe y. How to predict v? The "best linear predictor" (BLP) of some random term v is defined to be $\widetilde{v}$ for which $E\{(v-\widetilde{v})^2\}$ is minimized under the constraint that $\widetilde{v}$ is linear in y. The solution to this minimization problem is

$$\widetilde{v} = E[v|y]; \tag{4}$$

is called "the best linear predictor" of v. In general if v is vector then best linear prediction corresponds to minimization of $E\{\|v-\widetilde{v}\|^2\}$ and the solution is $\widetilde{v} = E[v|y]$.

We will apply this logic to our problem to estimate the BLUP of β and b. We want to minimize

$$E[\|(X\widetilde{\beta} + Z\widetilde{b}) - (X\beta + Zb)\|^2]$$

subject to the solution being unbiased,

$$E(X\widetilde{\beta} + Z\widetilde{b}) = E(X\beta + Zb).$$

It can be shown that the solutions are:

$$BLUP(\beta): \qquad \widetilde{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y \tag{5}$$

$$BLUP(b): \qquad \widetilde{b} = E[b|Y] = GZ^T\Sigma^{-1}(Y - X\widetilde{\beta}); \tag{6}$$

here $\Sigma = \mathrm{diag}(\Sigma_1, \ldots, \Sigma_n)$ is the $N \times N$ covariance matrix of Y.

Remark: The solution for β is the same as the GLS. The two solutions coincide with the maximizers of the joint likelihood discussed earlier. The earlier justification (Henderson's justification) makes distributional assumptions. The latter one does not make any distributional assumptions - it predicts the random parameters by *best prediction* using conditional expectation. However, the analytical expression of $\widetilde{b}$ is based on normality.

In practice one uses the plug-in estimators for Σ and $\widetilde{\beta}$, denoted by , $\widehat{\Sigma}$ and $\widehat{\beta}$, respectively.

Standard error estimation. From the BLUP estimation of β, it follows that

$$\mathrm{cov}\widetilde{\beta} = (X^T\Sigma^{-1}X)^{-1}$$

and so the standard errors of the components of β are determined based on the diagonal of this covariance matrix.

How to calculate standard errors for $\widetilde{b}$? Notice that to begin with b is random vector, and thus instead of using the typical idea to calculate $\mathrm{cov}\widetilde{b}$ we calculate $\mathrm{cov}\{b-\widetilde{b}\}$. More formally, we calculate the precision of the BLUP estimates as:

$$\begin{aligned}\mathrm{cov}\begin{bmatrix}\widetilde{\beta}-\beta\\ \widetilde{b}-b\end{bmatrix} &= \mathrm{cov}\begin{bmatrix}\widetilde{\beta}\\ \widetilde{b}-b\end{bmatrix} = \mathrm{cov}\left\{\begin{bmatrix}\widetilde{\beta}\\ \widetilde{b}\end{bmatrix}-\begin{bmatrix}0 & 0\\ 0 & I_q\end{bmatrix}\begin{bmatrix}\beta\\ b\end{bmatrix}\right\}\\ &= \mathrm{cov}(\widetilde{\theta}-M\theta)\mathrm{cov}(HC^TR^{-1}C\theta+HC^TR^{-1}e-M\theta)\\ &= H\end{aligned}$$

where $H=(C^TR^{-1}C+B)^{-1}$, $\theta=(\beta^T,b^T)^T$ and M matrix is like B, except the right-most lower block diagonal (analogous to G^{-1}) is I_q. Here we used the analytical solution of $\widetilde{\theta}=(C^TR^{-1}C+B)^{-1}C^TR^{-1}Y$ and the representation of Y as $Y=C\theta+e$. Also we used the fact that b and e are assumed independent.

The standard error in estimation of the BLUP for b is obtained based on the formula above. In practice we use plug-in estimators of G, $\widehat{G}$ based on $\widehat{D}$ and of R, $\widehat{R}$.

The BLUP estimator for b_i is

$$\widetilde{b}_i = DZ_i^T\Sigma_i^{-1}(Y_i-X_i\widehat{\beta}),$$

where $\widehat{\beta}$ is the GLS estimator of β. Notice Σ_i, D are known up to the set of parameters ω. When the covariance parameters estimates are plugged in the above estimator, the unbiasedness and smallest variance properties hold approximately.

The estimator

$$\widehat{b}_i = \widehat{D}Z_i^T\widehat{\Sigma}_i^{-1}(Y_i-X_i\widehat{\beta})$$

is called the empirical/approximate BLUP, also the Empirical Bayes estimator for b_i.

Prediction of individual trajectories Once the subject effects are estimated, then the individual mean can be estimated using the conditional expectation $E[Y_i|b_i]$; the estimator for this is

$$\widehat{Y_i} = \widehat{E[Y_i|b_i]} = X_i\widehat{\beta} + Z_i\widehat{D}Z_i^T\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta})$$

The subject's predicted response can be represented as a weighted average of the population mean profile and the observed subject's mean profile:

$$\widehat{Y_i} = (\widehat{R}_i\widehat{\Sigma}_i^{-1})X_i\widehat{\beta} + (I - \widehat{R}_i\widehat{\Sigma}_i^{-1})Y_i$$

Discussion: the predicted response is shrunk towards the population average profile, where the amount of 'shrinkage' depends on the relative magnitude of R_i and Σ_i, or their estimates

In class exercise: To gain more insight, consider the following example $Y_{ij} = \mu + b_i + e_{ij}$ for $b_i \sim$ iid $N(0, D)$ and $e_{ij} \sim iid\ N(0, \sigma^2)$, where D is scalar here.

i Estimate the population mean parameter μ.

ii Find the predicted b_i. Hint: $\Sigma_i^{-1} = \frac{1}{\sigma^2}(I_{m_i} - \frac{D}{\sigma^2 + m_i D} J_{n_i})$, where $var(Y_i) = \Sigma_i$ and J_{m_i} is the $m_i \times m_i$ matrix of ones.

iii Determine the individual predicted response

iv Discuss the effect of large/small D ('among units variance') in magnitude compared to σ^2.

Comparing nested models for the covariance: testing whether an effect is random

Motivation. Recall the six cities air pollution example, where FEV1 measurements are obtained for 299 girls observed at repeated occasions. The following model is assumed

$$\log(FEV1)_{ij} \quad = \quad (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})Age_{ij} + \beta_2 \log(Ht_{ij}) + e_{ij};$$

where $e_{ij} \sim N(0, \sigma^2)$ and

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{bmatrix} \right).$$

From the modeling perspective this covariance structure is appealing, because the number of covariance parameters is: 4. In general, for a model with q random effects, which assumes conditional independence, there are $q \times (q+1)/2 + 1$ covariance parameters.

For most models it is sufficient to assume random intercept and random slope (like above) ! In our model, it is assumed implicitly that Age_{ij} has a random effect, and not fixed. The question we are interested to study is whether Age has random effect or just fixed? Statistically we formulate the hypothesis that the subjects specific slopes are not random by the following hypothesis testing

$$H_0 : D_{22} = 0 \qquad \text{versus} \qquad H_1 : D_{22} \neq 0;$$

notice that the null hypothesis also imply that $D_{12} = 0$.

In general one is interested to *test whether there are q correlated random effects versus $(q-1)$ correlated random effects*. Assume the model is $Y_i = X_i\beta + Z_i b_i + e_i$, where $b_i \sim iidN_q(0, D)$ and

$$D = \begin{bmatrix} D_{11} & D_{12} \\ D_{12}^T & D_{22} \end{bmatrix},$$

where D_{11} is $(q-1) \times (q-1)$ matrix, D_{12} is $(q-1) \times 1$ matrix and $D_{22} \geq 0$. Of interest is to test the hypothesis testing

$$H_0 : D_{22} = 0, D_{12} = 0, D_{11} \text{ is positive definite} \qquad \text{versus} \qquad H_1 : D \text{ is positive definite.}$$

In many settings likelihood ratio testing (LRT) is a valid test for comparing nested models. Recall the LRT

$$T_{LRT} = 2\log \widehat{L}_{H_1} - 2\log \widehat{L}_{H_0};$$

the classical asymptotic null distribution of the LRT is χ^2_r, where r is the difference between the number of parameters specified by the alternative and the ones needed by the null. However standard theory used to develop the asymptotic null distribution of the test is chi-square is not valid. (Why?)

It turns out that the asymptotic null distribution of LRT is $0.5\chi^2_{q-1} + 0.5\chi^2_q$.

Brief intuition: Any variance quantity is non-negative; hence testing that a variance parameter equals zero means that we test for a value that is at the boundary of the parameter space. The classical theoretical arguments are no longer valid. If the usual null distribution is used the resulting p-value will be overestimated. Thus, in general, ignoring this problem can lead to selection of model for covariance that is too simple.

Final Remarks

- The REML likelihood provides a measure of the goodness of fit of an assumed model for the covariance. A standard approach for comparing two nested models is via the LRT with the correct asymptotic null distribution.

- For non-standard covariance comparisons or when the models are not nested, they can be compared in terms of information criteria (AIC, cAIC, BIC) that effectively penalize the complexity of the model assumed.

Here too, if the assumed covariance has been mis-specified, we can correct the standard errors by using 'empirical' or so-called 'robust' variances. The empirical or so-called 'robust variance' of $\widehat{\beta}$ is obtained by using $\widehat{V}_i = (Y_i - X_i\widehat{\beta})(Y_i - X_i\widehat{\beta})^T$ as an estimate of $var(Y_i)$. Recall that we derived the empirical variance of $\widehat{\beta}$ some time ago.

This empirical variance estimator is also known as the 'sandwich estimator'. The remarkable thing about the empirical estimator of $var(\widehat{\beta})$ is that it is a consistent estimator of the variance even when the model for the covariance matrix has been misspecified. That is, in large samples the empirical variance estimator yields correct standard errors.

In general, its use should be confined to cases where the number of subjects/objects n is relatively large and the number of repeated measurements m is relatively small. The empirical variance estimator may not be appropriate when there is severe imbalance in the data.

Bad Data.... Good Data