Week 6: Linear Mixed Models

MATH-516 Applied Statistics

Linda Mhalla

2024-03-27

Introduction

The goal today is to show how the linear regression model can be adapted to account for dependence between observations

- Consider a random vector $\mathbf{Y}$ of dimension n
  - such a vector would usually comprise repeated measures on an individual, or even observations from a group of individuals
- When independence fails, the estimated standard errors of the coefficients of the linear model are too small $\Rightarrow$ reject the null hypothesis more often then we should if the null is true (inflated Type I error, false positives)
- Need to account for within-group correlations, i.e., model a covariance matrix for observations within the same group (or within the same individual in the case of repeated measures)

Introduction

Longitudinal studies on independent subjects

- Measurements are taken from the same individuals, usually over time
    - These data are termed *repeated measures* or *longitudinal data*, but econometricians use the vocable *panel data*
    - The individuals are **independent** from one another; however, measurements from the same subject are not independent

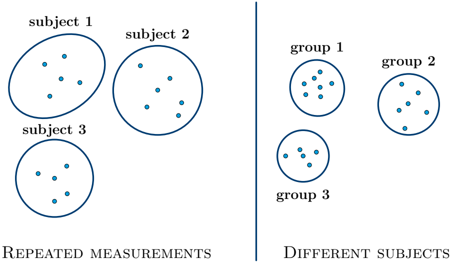
Studies on subjects that are not independent

- Subjects are sampled within a group
    - subjects sampled from the same household
    - subjects sampled from within several businesses
    - subjects sampled within schools, hospitals, etc

Introduction

We can always consider correlated data as grouped data, where there is within-group correlation

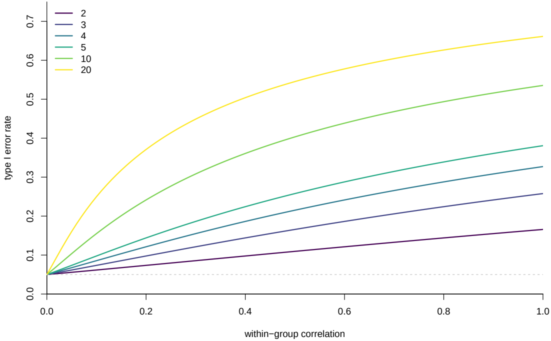
- In longitudinal data, we have several records for each individual
- In other examples, the groups could be households, schools, hospitals, businesses, etc.



One dot equals one line in the data file

What happens if we ignore within-group correlation?

Suppose that we have grouped data and we perform a one-sample t-test with level $\alpha = 5\%$



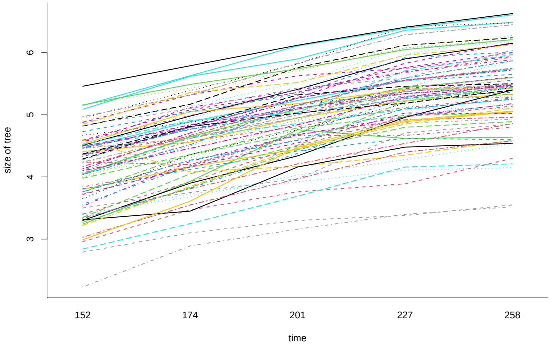
Type I error probability increases with correlation, as well as with the number of samples within each group ⇒ statistical inference is typically no longer valid ...

Example: Tree growth

- log-size (log-height+2*log-diameter) of 79 (Sitka spruce) trees measured repeatedly in about 1-month intervals
  - each tree measured 5-times
  - 54 trees grown in ozone-enriched environment (`treat=1`) and 25 were control

```
  size time tree treat
1 4.51  152    1 ozone
2 4.98  174    1 ozone
3 5.41  201    1 ozone
4 5.90  227    1 ozone
5 6.15  258    1 ozone
6 4.24  152    2 ozone
```

Example: Spaghetti plot



- Spaghetti plot that shows 79 curves (one for each tree)
- The size seems to increase with time, on average. The increase could be linear!

Section 1

Linear model with correlated errors

Notations

- Suppose that we collect observations from m groups such that:
  - There are n_i observations within group i $(i = 1, \dots, m)$
  - Any two observations from the same group are possibly correlated
  - Any two observations from different groups are assumed independent
- Groups can be formed in several ways:
  - Several measures can be taken from the same subject (repeated measures) and each individual forms a group
  - A group could also consist of individuals from the same school, department, or family
- We use the index i to indicate the group, and j to indicate an observation within a group
  - $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$ the outcome variable for group i
  - $\mathbf{X}_{ij} = (1, \mathrm{X}_{ij1}, \dots, \mathrm{X}_{ijp})^\top$ the set of p explanatory variables for observation j in group i

Linear model with correlated errors

The linear regression model is

$$Y_{ij} = \beta_0 + \beta_1 \mathrm{X}_{ij1} + \cdots + \beta_p \mathrm{X}_{ijp} + \varepsilon_{ij}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where ε_{ij} is the error term for observation j in group i

We assume that $\mathbb{E}(\varepsilon_{ij} \mid \mathbf{X}_{ij}) = 0$ and therefore

$$\mathbb{E}(Y_{ij} \mid \mathbf{X}_i) = \beta_0 + \beta_1 \mathrm{X}_{ij1} + \cdots + \beta_p \mathrm{X}_{ijp}$$

However, we no longer assume that the error terms are independent, i.e., ε and hence $\mathbf{Y}$ (when $\mathbf{X}$ is fixed) are assumed correlated

Linear model with correlated errors

- We assume the groups are independent from one another, so $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ if $i \neq i'$
- We model the **within-group** correlation by assuming that the covariance matrix of $\mathbf{Y}$ for group i is

  $$\mathrm{Cov}(\mathbf{Y}_i \mid \mathbf{X}_i) = \mathrm{Cov}(\varepsilon_i \mid \mathbf{X}_i) = \Sigma_i,$$

  where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$ is the vector of errors for group i
- Assuming data re ordered by group, the full covariance matrix is block-diagonal

$$\mathrm{Cov}(\mathbf{Y}) = \begin{pmatrix} \Sigma_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \Sigma_2 & \cdots & \mathbf{O} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \Sigma_m \end{pmatrix}$$

Covariance structures

- Compound symmetry: observations within a group are interchangeable

$$\Sigma_i = \begin{pmatrix} \sigma^2+\tau & \tau & \tau & \tau & \tau \\ \tau & \sigma^2+\tau & \tau & \tau & \tau \\ \tau & \tau & \sigma^2+\tau & \tau & \tau \\ \tau & \tau & \tau & \sigma^2+\tau & \tau \\ \tau & \tau & \tau & \tau & \sigma^2+\tau \end{pmatrix}$$

- Auto-regressive structure `AR(1)`: magnitude of correlation depends on amount of time between observations

$$\mathbf{R}_i = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \text{ with } \Sigma_i = \sigma^2\mathbf{R}_i$$

Covariance structures

- Heterogeneous AR structure `ARH(1)`: same correlation matrix as `AR(1)` but covariance matrix

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 & \sigma_1\sigma_5\rho^4 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 & \sigma_2\sigma_5\rho^3 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho^2 \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 & \sigma_4\sigma_5\rho \\ \sigma_5\sigma_1\rho^4 & \sigma_5\sigma_2\rho^3 & \sigma_5\sigma_3\rho^2 & \sigma_5\sigma_2\rho & \sigma_5^2 \end{pmatrix}$$

- Unstructured covariance matrix ...

Tree growth example: we get 15 parameters under the unstructured model, compared to two parameters for the compound symmetry and the `AR(1)` covariance models, and to six for the `ARH(1)` covariance model

Comparing covariance structures

- Many models are nested so use formal likelihood ratio tests whenever possible for comparisons
  - e.g., independence $\prec$ AR(1) $\prec$ ARH(1) $\prec$ unstructured
- Using AIC or BIC to compare models is valid **provided** the mean model includes the **same** variables
- When inference relies on ML (rather than REML), AIC and BIC can be used to compare models with different variables for the mean

Details on inference to follow ...

Section 2

Linear Mixed Effet Models

Introduction: Tree growth example

- So far, we have only accounted for group structure by modelling the within-group correlation
- We may also want to include a **group/individual effect** in the mean model, e.g., a different intercept (and/or slope) for each group/individual

Suppose that the tree growth is approximately linear

$$Y_{ij} = \beta_{i1} + \beta_{i2} t_{ij} + \beta_3 \texttt{Treat}_i + \epsilon_{ij}, \quad \epsilon_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \Sigma_i)$$

The effect of `Treat` is not identifiable ...

- Collinearity issues: a variable fixed in time is perfectly collinear with the group/individual variable
- We cannot have a fixed effect for each tree while simultaneously including variables that are fixed in time (ozone treatment)

Introduction: A two-stage approach

Stage 1: Separate linear models *for each tree* i, i.e., assume the growth of each tree is approximately linear with tree-specific intercepts and slopes:

$$Y_{ij} = \beta_{i1} + \beta_{i2} t_{ij} + \epsilon_{ij}, \quad \epsilon_i \sim \mathcal{N}_{n_i}\left(\mathbf{0}, \Sigma_i\right)$$

With $\mathbf{Y}_i = \left(Y_{i1}, \ldots, Y_{in_i}\right)^T, \beta_i = \left(\beta_{i1}, \beta_{i2}\right)^T$ and

$$\mathbf{Z}_i = \begin{pmatrix} 1 & 1 & \ldots & 1 & \ldots & 1 \\ t_{i1} & t_{i2} & \ldots & t_{ij} & \ldots & t_{in_i} \end{pmatrix}^T$$

and with distributional assumption (normality), we can write this as

$$\mathbf{Y}_i = \mathbf{Z}_i \beta_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}_{n_i}\left(\mathbf{0}, \Sigma_i\right)$$

Introduction: A two-stage approach

Stage 2: Regression model for the coefficients $\beta_i = (\beta_{i1}, \beta_{i2})^T$

$$\beta_{i1} = \beta_1 + \beta_2 \texttt{Treat}_i + b_{i1}, \quad \beta_{i2} = \beta_3 + \beta_4 \texttt{Treat}_i + b_{i2},$$

i.e., β_{i1} and β_{i2} are tree-specific intercepts and slopes depending on ozone treatment $\texttt{Treat}_i$

With $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T$, $\mathbf{b}_i = (b_{i1}, b_{i2})^T \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D})$, and

$$\mathbf{K}_i = \begin{pmatrix} 1 & \texttt{Treat}_i & 0 & 0 \\ 0 & 0 & 1 & \texttt{Treat}_i \end{pmatrix}$$

and with distributional assumption, the second-stage model is

$$\beta_i = \mathbf{K}_i \beta + \mathbf{b}_i, \quad \mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D})$$

⇒ systematic differences between treated and control trees

⇒ individual intercepts/slopes that are normally distributed around their treatment means

Introduction: A two-stage approach

The resulting combined model is

$$\begin{aligned}\mathbf{Y}_i &= \mathbf{Z}_i\mathbf{K}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, \quad \mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D}) \perp\!\!\!\perp \epsilon_i \sim \mathcal{N}\mathbf{0}, \Sigma_i) \\ &= \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i\end{aligned}$$

Disadvantages:

- We often have few observations per group/individual to estimate β_i
- Uncertainty assessment is tricky as plug-in estimate $\widehat{\beta}_i$ replaces β_i in second-stage model

Conclusion:

Combining both models in one model seems more adequate → the combined model is the **linear mixed model**

- β are called fixed effects: population effects
- The residuals $\mathbf{b}_i$ are normally distributed and are termed the **random effects**: group/individual-specific effects

The linear mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, , \quad \mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D}) \perp\!\!\!\perp \epsilon_i \sim \mathcal{N}(0, \Sigma_i)$$

- $\mathbf{X}_i$ can include time-constant and time-varying variables (interaction between time and covariates too, e.g., tree growth)
- $\mathbf{X}_i$ should include the covariates in $\mathbf{Z}_i$, as $\mathbb{E}(\mathbf{b}_i) = \mathbf{0}$

The main characteristic of the **linear mixed model** is to allow certain variables to have random effects, i.e., to have parameters that vary from one group/individual to another

- This captures heterogeneity between groups/individuals
- While each group is allowed an individual effect, the overall average of these effects is zero

Marginal versus conditional view

In this model, we still have the so-called marginal mean of Y_{ij}

$$\mathbb{E}(\mathbf{Y}_i) = \mathbf{X}_i\beta$$

- At the population level, the mean of $\mathbf{Y}_i$ is only a function of the fixed effects

We also have the conditional mean of $\mathbf{Y}_i$, which depends on the group-specific effect

$$\mathbb{E}(\mathbf{Y}_i \mid \mathbf{b}_i) = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i$$

- The random effects are group-specific mean effects
- The mean of $\mathbf{Y}_i$ is a function of population and group-specific effects

Marginal versus conditional view

Since $\mathbf{b}_i$ are random terms, they introduce a within-group correlation in the model

- The marginal variance is

$$\mathrm{Cov}(\mathbf{Y}_i) = \Sigma_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top$$

→ a sum of deviations of groups from the population average + deviations of observations from their group's mean trend

- The conditional variance is

$$\mathrm{Cov}(\mathbf{Y}_i \mid \mathbf{b}_i) = \Sigma_i$$

Random intercept model

A very common special case is the random intercept model

$$Y_{ij} = \mathbf{x}_{ij}^T \beta + b_i + \epsilon_{ij},$$

where $\mathbf{x}_{ij}$ is the covariate vector for the j th measurement of the i th group/individual

For the tree growth example, let's assume the model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{ij}$$

with $b_i \sim \mathcal{N}(0, d)$ ⇒ only the intercept varies between the trees

Assuming independent and homogeneous errors $\epsilon_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 \mathbf{I})$, we get

$$\text{Cov}\left(Y_{ij}, Y_{ik}\right) = d + \sigma^2 I(j = k)$$
$$\Rightarrow \text{Corr}\left(Y_{ij}, Y_{ik}\right) = \frac{d}{d + \sigma^2} = \rho, \quad j \neq k$$

⇒ Compound symmetry correlation structure

Example: Tree growth

Consider a simple model, where only the intercept is random:

$$\mathbb{E}[Y_{ki} \mid X_{ki} = t_{ki}, b_k] = (\beta_0 + b_k) + \beta_1 t_{ki}$$

and the corresponding fixed-effect-only model `y ~ tree+time`.

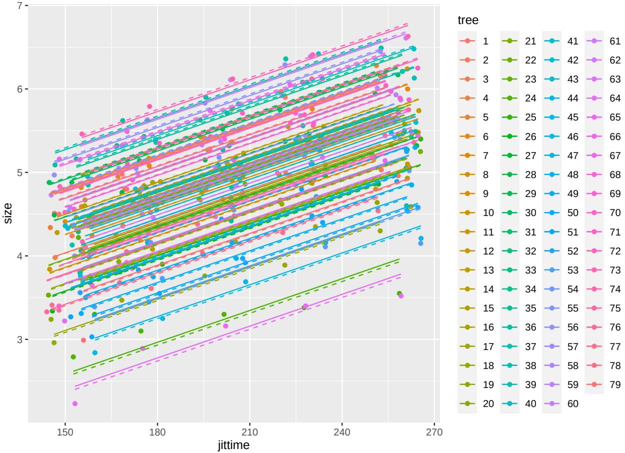


Figure: Tree growth data and lines by fixed−effect−only model (dashed) and random intercept model (solid).

Mixed linear models in R

To fit linear mixed models in R, one can use

- function `lme` in the package nlme (see Pinheiro and Bates, 2000)
  - structure is similar to `lm` but with argument `random`
    - `random = ~ 1 |subject` : random intercepts for each group/subject
    - `random = ~ 1 + time |subject` : random intercepts and slopes for each group/subject
    - multilevel models with several nested random effects (see this link for details on multilevel models):
      `random = ~ 1+time | hospital/subject`
- function `lmer` in the package lme4 (see Bates et al, 2015)
  - includes GLM, via the glmer function
  - does not implement heteroscedasticity of residuals

For a larger class of linear mixed models including, e.g., smooth terms, see

- functions `gam` and `bam` (for large data) in the package `mgcv`

Section 3

Estimation of the LMM

The marginal model

Estimation is usually based on the marginal model

The linear mixed model

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i \\ \mathbf{b}_i \sim \mathcal{N}\left(\mathbf{0}, \mathbf{D}\right) \\ \epsilon_i \sim \mathcal{N}\left(\mathbf{0}_{n_i}, \Sigma_i\right) \\ \mathbf{b}_1, \dots, \mathbf{b}_m, \epsilon_1, \dots, \epsilon_m \text{ independent} \end{cases}$$

implies the marginal model

$$\mathbf{Y}_i \sim \mathcal{N}\left(\mathbf{X}_i\beta, \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \Sigma_i\right), \text{ for } i = 1, \dots, m$$

Estimation of fixed effects

- Let α denote the vector of parameters of $\mathbf{V}_i$, i.e., the elements in $\mathbf{D}$ and Σ_i (e.g., σ^2 if $\Sigma_i = \sigma^2 \mathbf{I}_{n_i}$)
- Let $\theta = (\beta, \alpha)$

Then, the marginal log-likelihood (log-likelihood of the marginal model) is

$$\ell_{ML}(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}(\alpha)| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{V}(\alpha)^{-1}(\mathbf{y} - \mathbf{X}\beta)$$

Let's assume that α is known and focus on estimation of the fixed effects:

$$\frac{\partial}{\partial\beta}\ell_{ML}(\theta) = \mathbf{X}^\top \mathbf{V}(\alpha)^{-1}(\mathbf{y} - \mathbf{X}\beta) \overset{!}{=} \mathbf{0}$$
$$\Rightarrow \hat{\beta}_{ML}(\alpha) = \left\{\mathbf{X}^\top \mathbf{V}(\alpha)^{-1}\mathbf{X}\right\}^{-1} \mathbf{X}^T \mathbf{V}(\alpha)^{-1}\mathbf{y}$$

$\Rightarrow \hat{\beta}_{ML}(\alpha)$ is a weighted least square estimator

Estimation of variance parameters

- Substituting $\widehat{\beta}_{ML}(\alpha)$ into the marginal log-likelihood gives the profile log-likelihood $\ell_{ML}(\widehat{\beta}_{ML}(\alpha), \alpha)$

$\Rightarrow$ maximize $\ell_{ML}(\widehat{\beta}_{ML}(\alpha), \alpha)$ numerically to obtain the ML estimator $\hat{\alpha}_{ML}$

- But, ML estimators of variance are known to be biased (downwards)

$\Rightarrow$ estimation by **restricted maximum likelihood (REML)**

Intuition of REML: Instead of working with $\mathbf{Y}$, work with its linear transformation $\mathbf{U} = \mathbf{A}^\top \mathbf{Y}$ s.t. $\mathbb{E}(\mathbf{U}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{U}) = \sigma^2 \mathbf{A}^\top \mathbf{A}$. Then, maximize the likelihood based on $\mathbf{U}$ (does not involve the mean)

The matrix $\mathbf{A}$ is chosen orthogonal to design matrix $\mathbf{X}$ $\Rightarrow$ for two models with different design matrices, we use different $\mathbf{U}$s and their REML likelihoods are not comparable

Interpretation of variance components

```
library(mixedup)
library(knitr)
library(lme4)

mm1 <- lmer(size~ time*treat + (1|tree), data=Sitka, REML=TRUE)
# summary(mm1)
```

Table 1: Estimated variance of random effects

group	effect	variance	sd	sd_2.5	sd_97.5	var_prop
tree	Intercept	0.370	0.608	0.516	0.710	0.908
Residual	NA	0.038	0.194	0.179	0.209	0.092

The estimated sd of the tree effect tells us how much, on average, size differs as we move from a tree to another

- the intra-group/intra-individual correlation: $0.37/(0.37+0.038) \approx$ 0.907

Prediction of random effects

The terms $\mathbf{b}_i$ are random variables that can be **predicted** relying on the conditional model (and not the marginal)

$$\mathbf{Y}_i|\mathbf{b}_i \sim \mathcal{N}_{n_i}(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i, \Sigma_i)$$

Usually, $\hat{\mathbf{b}}_i(\theta) = \mathbb{E}(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i)$, with (hidden) parameters α and β replaced by their ML or REML estimates

Section 4

Inference for the LMM

Testing for the fixed effects

- Wald test relying on the asymptotic normality of $\hat{\beta}$

$$(\hat{\beta} - \beta) \dot{\sim} \mathcal{N}(\mathbf{0}, (\mathbf{X}^\top \mathbf{V}(\hat{\alpha})^{-1} \mathbf{X})^{-1})$$

- LRT for nested models
  - Restricted likelihoods are not comparable when fixed effects differ
  - LRT can only be used with ML estimation (and not with REML)

Example: Tree growth

```r
library(nlme)

mm1 <- lme(size ~ time, random = ~ 1|tree, data=Sitka,
           method="ML")
m1  <- lme(size ~ time * treat, random = ~ 1|tree, data=Sitka,
           method="ML")

anova(m1,mm1)
```

```
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m1      1  6 142.1890 166.0623 -65.09451                        
mm1     2  4 154.6453 170.5608 -73.32263 1 vs 2 16.45623   3e-04
```

- ozone treatment is significant

Testing for the random effects

Testing the need for a random effect is equivalent to testing that its variance is null

⇒ MLE regularity assumptions are typically not met: Under the null, the parameter does not lie in the interior but on the boundary of the parameter space, as 0 is on the boundary of $[0, \infty)$

Denote

$$\mathrm{Cov}\left(\mathbf{b}_i\right) = \mathbf{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$$

Consider three possible models:

- M_0 : no random effects $\left(\mathbf{b}_i \equiv \mathbf{0}\right), d_{11} = d_{12} = d_{22} = 0$
- M_1 : only a random intercept $\left(b_{2i} \equiv 0\right), d_{12} = d_{22} = 0$
- M_2 : (correlated) random intercept and slope

We can compare M_2 and M_1 by testing for $H_{0,1} : d_{12} = d_{22} = 0$ and M_1 and M_0 by testing for $H_{0,2} : d_{11} = 0$

Testing for the random effects

- Testing for $H_{0,2} : d_{11} = 0$
  - the LRT statistic is not asymptotically χ^2_1 distributed but is rather a mixture of a point mass at 0 (half of the time) and a χ^2_1 distribution, under the null (recall testing between a negative binomial and a Poisson) $\rightarrow$ divide the p-value by two
  - For $\Sigma_i = \sigma^2\mathbf{I}$, an exact distribution is available (Crainiceanu and Ruppert, 2004); see the `R` package `RLRsim`
- Testing for $H_{0,1} : d_{12} = d_{22} = 0$
  - the LRT statistic is not asymptotically χ^2_2 distributed but is rather a mixture of a χ^2_1 (half of the time) and a χ^2_2 distribution, under the null

Example: Tree growth

```r
mm1 <- lme(size ~ time, random = ~ 1|tree, data=Sitka,
           method="REML")
mm2 <- lme(size ~ time, random = ~ time|tree, data=Sitka,
           method="REML")

anova(mm1, mm2)
```

```
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mm1     1  4 172.7768 188.6720 -82.38840                        
mm2     2  6 136.9669 160.8098 -62.48344 1 vs 2 39.80992  <.0001
```

```r
t.stat <- 39.80992
p.value <- 0.5 * (1-pchisq(t.stat,1)) + 0.5 * (1-pchisq(t.stat,2))

p.value
```

```
[1] 1.273288e-09
```