# 36 线性混合模型

## 36.1 介绍

在基本回归分析模型中， 假定回归误差项独立同分布， 方差相等， 还经常假定误差项服从正态分布。 在实际应用回归分析建模时， 还经常遇到模型误差项方差不相等或者误差项之间不独立的情形。 比如， 如果每个观测来自于一个群体的平均值， 设群体中的个体方差相同， 则每个观测的误差方差正比于群体中个体的个数的倒数， 不等于常数。 又比如， 医学研究中每个受试者有多次随访的数值型观测值， 则每个受试者的各次测量值是相关的， 不同受试者之间可以认为是独立的， 这样模型误差不是相互独立的。

考虑如下的线性模型：

$$\begin{aligned}\boldsymbol{Y} =& X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon} \sim& \mathrm{N}(\boldsymbol{0}, \sigma^2 W^{-1}).\end{aligned}$$

其中$\boldsymbol{X}$为$n \times p$非随机的自变量矩阵， $\boldsymbol{\beta}$是$p$维回归系数向量， $\boldsymbol{\varepsilon}$为$n$维随机误差向量， $W$已知，$\sigma^2$未知。

因为标准的线性模型要求$W = I$， 所以这个模型是线性模型的推广。

考虑这个模型的估计问题。 设存在矩阵$A$使得$W^{-1} = AA^T$， 则

$$A^{-1}\boldsymbol{Y} = A^{-1}X\boldsymbol{\beta} + A^{-1}\boldsymbol{\varepsilon},$$

令$\boldsymbol{Y}^* = A^{-1}\boldsymbol{Y}$, $X^* = A^{-1}X$, $\boldsymbol{\varepsilon}^* = A^{-1}\boldsymbol{\varepsilon}$， 则

$$\boldsymbol{Y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*,$$

其中$\mathrm{Var}(\boldsymbol{\varepsilon}^*) = \sigma^2 I$。 在矩阵可逆条件下$\boldsymbol{\beta}$的最小二乘估计为

$$\hat{\boldsymbol{\beta}} = (X^{*T}X^*)^{-1}X^{*T}\boldsymbol{Y}^* = (X^TWX)^{-1}X^TW\boldsymbol{Y}.$$

这个公式称为加权最小二乘公式。

模型可以进一步推广。 考虑

$$\begin{aligned}\boldsymbol{Y} =& X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon} \sim& \mathrm{N}(\boldsymbol{0}, \sigma^2 W^{-1}),\\ \boldsymbol{\alpha} \sim& \mathrm{N}(\boldsymbol{0}, \Sigma).\end{aligned}$$

其中$\boldsymbol{Y}$为可观测的$n$维因变量向量， $X_{n\times p}$, $Z_{n\times q}$已知， $\boldsymbol{\beta}_{p\times 1}$为未知的非随机回归系数向量，称为**固定效应**； $\boldsymbol{\alpha}_{q\times 1}$为未知的随机向量， 称为**随机效应**， $\Sigma$未知； $\boldsymbol{\varepsilon}$为未知的随机误差向量，$\sigma^2$未知，$W$已知。 称此模型为**线性混合模型**。

在估计模型时， $\Sigma$必须是对称非负定阵， 一般有一定的参数结构，所以一般会分解为

$$\Sigma = \sigma^2 \Lambda_\theta \Lambda_\theta^T,$$

其中$\theta$为待估参数。

因为随机效应$\boldsymbol{\alpha}$未知但一般不需要估计， 所以可以将其作用混合到因变量$\boldsymbol{Y}$的方差结构中， 得到

$$\begin{aligned} E\boldsymbol{Y} =& X\boldsymbol{\beta}, \\ \mathrm{Var}(\boldsymbol{Y}) =& \sigma^2 W^{-1} + Z\Sigma Z^T = \sigma^2 (W^{-1} + (Z\Lambda_\theta)(Z\Lambda_\theta)^T). \end{aligned}$$

线性混合效应模型适用的数据包括：

- 集簇数据，比如， 同一学科的不同授课教师的学生成绩；
- 重复量测，比如多个受试者每人都多次测量某一指标；
- 纵向数据，比如多个病人的多次不同时间的随访观测；
- 多元观测，比如多个病人每人都有多个生理指标。

本章理论和例子参考了(Andrzej Galecki 2013)。

本章主要使用nmleU包的armd数据。 ARMD（年龄相关性黄斑变性）是一种老年人眼科疾病，会使得患病的人逐渐失明。 为了研究某种新药的疗效， 收录了240位病人， 随机分为治疗组与对照组， 并在入组时以及4、12、24、52周时对每位病人测量视力指标。 这样的数据称为纵向数据， 每位病人都测量了5个不同时间的因变量值， 但这5个时间是共同的。 有些病人的随访有缺失， 并且有些是在中间缺失的。 每个病人的各次视力指标都与入组时的基础测量值有关。

```
data(armd, package="nlmeU")
str(armd)
```

```
## 'data.frame':    867 obs. of  8 variables:
##  $ subject : Factor w/ 234 levels "1","2","3","4",..: 1 1 2 2 2 2 3 3 3 4 ...
##  $ treat.f : Factor w/ 2 levels "Placebo","Active": 2 2 2 2 2 2 1 1 1 1 ...
##  $ visual0 : int  59 59 65 65 65 65 40 40 40 67 ...
##  $ miss.pat: Factor w/ 8 levels "----","---X",..: 4 4 1 1 1 1 2 2 2 1 ...
##  $ time.f  : Ord.factor w/ 4 levels "4wks"<"12wks"<..: 1 2 1 2 3 4 1 2 3 1 ...
##   ..- attr(*, "contrasts")= num [1:4, 1:3] -0.5222 -0.3023 0.0275 0.797 0.565 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:4] "4wks" "12wks" "24wks" "52wks"
##   .. .. ..$ : chr [1:3] ".L" ".Q" ".C"
##  $ time    : num  4 12 4 12 24 52 4 12 24 4 ...
##  $ visual  : int  55 45 70 65 65 55 40 37 17 64 ...
##  $ tp      : num  1 2 1 2 3 4 1 2 3 1 ...
```

# 36.2 加权线性回归

## 36.2.1 加权已知情形

对于某些回归问题的数据， 各个观测的因变量方差不相等。 设模型为：

$$\begin{aligned} y_i =&\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \\ &\varepsilon_i\text{相互独立，零均值，方差为}\frac{\sigma^2}{w_i},\ i = 1, 2, \ldots, n. \end{aligned}$$

这个模型违反了线性回归模型要求误差项方差相等的要求， 但可以通过简单的变换变成标准的线性回归问题。 令$y_i^* = \sqrt{w_i} y_i$, $x_{ij}^* = \sqrt{w_i} x_{ij}$, $\varepsilon_i^* = \sqrt{w_i}\varepsilon_i$， 则模型变成：

$$\begin{aligned} y_i^* =&\beta_0 + \beta_1 x_{i1}^* + \cdots + \beta_p x_{ip}^* + \varepsilon_i^*, \\ &\varepsilon_i^*\text{相互独立，零均值，方差为}\sigma^2,\ i = 1, 2, \ldots, n. \end{aligned}$$

于是最小二乘估计为：

$$\begin{aligned} &\min \sum_{i=1}^n \left[ y_i^* - \left( \beta_0 + \beta_1 x_{i1}^* + \cdots + \beta_p x_{ip}^* \right) \right]^2 \\ =&\min \sum_{i=1}^n w_i \left[ y_i - \left( \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \right) \right]^2, \end{aligned}$$

用向量记号表示的估计公式为

$$\hat{\boldsymbol{\beta}} = (X^T W X)^{-1} X^T W Y,$$

其中$W = \mathbf{diag}(w_1, \ldots, w_n)$。

常见的加权线性回归问题是每个观测都是某一类数据的平均值， 已知参与平均的个数但没有具体的原始数据。 在 `lm()` 函数中用 `weights=` 指定权重， 如果已知每个观测是 `n` 个观测的平均，可以指定 `weights=n` ， 用数据集中的变量 `n` 保存每个观测代表的原始观测个数。

考虑nlmeU包的armd数据框。 共有240为受试者分为处理组和对照组， 在第4、12、24、52周的安排了视力的随访测量， 这些随访的测量值可能有缺失。 将每个受试者的随访测量值进行平均， 利用视力的平均测量值比较处理组与对照组， 以基线的视力测量值为协变量， 进行加权回归时以参与平均的测量值个数作为权重（这是一种近似，因为有序列相关可能）。 程序如下：

```
data(armd, package="nlmeU")


dar <- armd |>
  group_by(subject) |>
  mutate(meanvis = mean(visual, na.rm=TRUE),
         nfu = sum(!is.na(visual))) |>
  filter(nfu > 0)
lm1.armd <- lm(meanvis ~ treat.f + visual0,
  weights = nfu, data=dar)
summary(lm1.armd)
```

```
##
## Call:
## lm(formula = meanvis ~ treat.f + visual0, data = dar, weights = nfu)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -52.506 -12.042   3.502  13.797  43.560
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.34194    1.31211   3.309 0.000975 ***
## treat.fActive -3.27536    0.66054  -4.959 8.55e-07 ***
## visual0        0.83199    0.02227  37.354  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19 on 864 degrees of freedom
## Multiple R-squared:  0.6233, Adjusted R-squared:  0.6224
## F-statistic: 714.8 on 2 and 864 DF,  p-value: < 2.2e-16
```

加权回归也可以用nlme包的 `gls()` 函数计算， `gls()` 可以进行误差项方差可变或者相关的广义最小二乘估计。 如：

```
library(nlme)
lm2.armd <- gls(meanvis ~ treat.f + visual0,
  weights = varFixed(value = ~ I(1/nfu)), data=dar)
summary(lm2.armd)
```

```
## Generalized least squares fit by REML
##   Model: meanvis ~ treat.f + visual0
##   Data: dar
##        AIC      BIC    logLik
##   6422.203 6441.249 -3207.102
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~I(1/nfu)
##
## Coefficients:
##                   Value Std.Error  t-value p-value
## (Intercept)    4.341935 1.3121127  3.30912   0.001
## treat.fActive -3.275356 0.6605417 -4.95859   0.000
## visual0        0.831994 0.0222731 37.35417   0.000
##
##  Correlation:
##               (Intr) trt.fA
## treat.fActive -0.260
## visual0       -0.938  0.024
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.7637685 -0.6338519  0.1843284  0.7262547  2.2928684
##
## Residual standard error: 18.99789
## Degrees of freedom: 867 total; 864 residual
```

其中 `weights()` 选项可以指定各种表示方差结构的函数， 这里的 `varFixed()` 函数指定一个自变量作为误差方差的倍数， 是权重的倒数。

## 36.2.2 方差依赖于均值的情形

比上述加权已知的情形略复杂一些的， 是$\varepsilon_i$的方差为如下函数：

$$\lambda_i = \sigma^2 \lambda(\mu_i),$$

其中$\mu_i = \boldsymbol{x}^i\boldsymbol{\beta}$， $\boldsymbol{x}^i$表示第$i$个观测的自变量。 这时， 可以用IRLS(Iteratively Re-Weighted Least Square)算法估计模型参数：

1. 给定$\boldsymbol{\beta}$的初值$\boldsymbol{\beta}^{(0)}$， 令迭代计数器$k = 0$;
2. 令$\mu_i^{(k)} = \boldsymbol{x}^i\boldsymbol{\beta}^{(k)}$， $\hat{\lambda}_i^{(k)} = \lambda(\mu_i^{(k)})$;
3. 取权重$w_i = 1/\hat{\lambda}_i^{(k)}$， 用加权最小二乘方法估计参数$\boldsymbol{\beta}$， 作为$\boldsymbol{\beta}^{(k+1)}$;
4. 令$k \leftarrow k + 1$;
5. 重复第2至第4步直到收敛。
6. 计算最终的权重，用加权最小二乘方法估计参数$\sigma^2$。

nmle包的 `gls()` 函数支持这样的方差结构， 并可以进行参数估计和推断。

# 36.3 线性混合模型

## 36.3.1 问题介绍

在医学研究等领域问题中， 经常收集多个受试者在多个不同时间的研究数据， 试图介绍某一因变量如何受到各种连续型以及分类型自变量的影响。 因为每人有沿时间收集的多个因变量值，所以这样的问题有些像是多元时间序列问题， 但是， 收集数据的时间点不一定是完全对齐的，不同受试者的时间点个数不一定相同， 所以经典的等时间间隔的多元时间序列模型如向量自回归无法使用。 另外， 时间序列模型也不擅长分析外生自变量的影响。

这种问题更适用于回归类的模型， 只不过要将回归模型中误差项独立同分布等方差假定放松，因为同一个受试者在不同时间的因变量值是相关的， 不同受试者的因变量值仍可以认为独立。在这样的模型中， 用来解释因变量的连续型和分类型自变量的作用称为**固定效应**， 个体之间的差别的影响称为**随机效应**， 这样既有固定效应又有随机效应的线性模型称为**线性混合模型**或者线性混合效应模型。

以nmleU包的ARMD数据为例， 这个数据中有240名ARMD（年龄相关性黄斑变性）， 随机分为治疗组与对照组， 并在入组时以及4、12、24、52周时对每位病人测量视力指标， 这样的数据称为纵向数据， 研究目的是考察治疗的有效性以及病变随时间的变换规律。 以入组后的4次视力测量值为因变量， 以处理效应、时间效应等为固定效应， 以不同病人作为随机效应， 可以用线性混合模型建模。

## 36.3.2 理论模型

设数据按某个分组变量分为$N$组， 不同组之间相互独立， 第$i$组有$n_i$个观测， 模型写成

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i.$$

模型(36.3)中：

- $\boldsymbol{X}_i$是由连续型自变量与因子自变量构造的回归设计阵， $\boldsymbol{\beta}$代表这些自变量的作用（固定效应）， 相当于经典的线性模型中的回归系数向量。
- $\boldsymbol{b}_i$是一个未知的随机向量， 代表第$\boldsymbol{i}$组特有的个体效应（随机效应）， 不同组的$\boldsymbol{b}_i$相互独立， $Z_i$为$n_i \times q$的设计阵， 保存与$q$个随机效应对应的自变量值， $Z_i\boldsymbol{b}_i$可以描述第$\boldsymbol{i}$组特有的相关性。 $Z_i$可以使用$\boldsymbol{X}_i$的部分列。
- $\boldsymbol{\varepsilon}_i$是第$\boldsymbol{i}$组的随机误差， 仍是分量之间独立的且在不同组之间独立的。

在模型中， 固定效应$\boldsymbol{\beta}$是需要估计和推断的， 而随机效应$\boldsymbol{b}_i$则一般不需要估计。 关于$\boldsymbol{b}_i$和$\boldsymbol{\varepsilon}_i$还进行如下的分布假定：

$$\boldsymbol{b}_i \sim \mathrm{N}_q(\boldsymbol{0}, \sigma^2 D), \quad \boldsymbol{\varepsilon}_i \sim \mathrm{N}_{n_i}(\boldsymbol{0}, \sigma^2 R_i), \quad \boldsymbol{b}_i\text{与}\boldsymbol{\varepsilon}_i\text{相互独立}.$$

不同组的$(\boldsymbol{b}_i, \boldsymbol{\varepsilon}_i)$之间也相互独立。

(36.3)是对第$\boldsymbol{i}$组的模型。 将所有组的$\boldsymbol{y}_i$合并成一个向量$\boldsymbol{y}$， 长度为$n = \sum_{i=1}^{N} n_i$， 将所有设计阵$\boldsymbol{X}_i$纵向合并成一个$n \times p$矩阵$\boldsymbol{X}$， 将所有随机效应$\boldsymbol{b}_i$合并成一个向量$\boldsymbol{b}$， 长度为$Nq$， 将所有对应于$\boldsymbol{b}_i$的设计阵$Z_i$按分块对角矩阵方式合并为一个$n \times (Nq)$矩阵$Z$， 将所有$\boldsymbol{\varepsilon}_i$合并成一个向量$\boldsymbol{\varepsilon}$， 模型可以统一写成

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\varepsilon},$$

其中

$$\boldsymbol{b} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \tilde{D}), \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 R),$$

$\boldsymbol{b}$和$\boldsymbol{\varepsilon}$相互独立， $\tilde{D}$和$R$为分块对角矩阵， $\tilde{D} = \mathrm{diag}(D, D, \ldots, D)$,
$R = \mathrm{diag}(R_1, \ldots, R_N)$。

随机效应和误差项的这种分块对角矩阵形式的方差阵结构， 是来源于假定不同组之间相互独立。 如果有两层分组， 内层分组嵌套在外层分组中， 这时可以引入外层与内层的两种随机效应， 在组间仍可假定相互独立， 随机效应和误差项的方差阵仍保持对角矩阵形式。 如果存在两种分组但分组之间存在交互则会违背这样的假定。

## 36.3.3 因变量分布

在每一组中， 若将$\boldsymbol{b}_i$看作已知， $\boldsymbol{y}_i$的条件分布服从多元正态分布， 有

$$\begin{aligned} E(\boldsymbol{y}_i|\boldsymbol{b}_i) \equiv \boldsymbol{\mu}_i &= X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \\ \mathrm{Var}(\boldsymbol{y}_i|\boldsymbol{b}_i) &= \sigma^2 R_i. \end{aligned}$$

一般形式的线性混合模型包含太多的未知参数， 无法估计， 所以需要对随机效应和误差项的方差结构进行简化， 假定其仅依赖于少数几个未知参数。 对于一个协方差阵$\boldsymbol{\Sigma}$， 可以将其分解为标准差与相关阵两部分：

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Xi}\boldsymbol{\Lambda},$$

其中$\boldsymbol{\Xi}$为$\boldsymbol{\Sigma}$对应的相关系数阵， $\boldsymbol{\Lambda}$是各分量的标准差组成的对角阵。 这样， 可以分别对标准差和相关系数进行参数化表示。

$\boldsymbol{y_i}$边缘分布为多元正态，

$$\begin{aligned} E(\boldsymbol{y}_i) =& X_i\boldsymbol{\beta}, \\ \mathrm{Var}(\boldsymbol{y}_i) =& \sigma^2\left(Z_i D Z_i^T + R_i\right). \end{aligned}$$

从给定$\boldsymbol{b_i}$的条件分布与边缘分布来看， 固定效应系数$\boldsymbol{\beta}$起到了相同的作用， 可以将其看成总体（所有受试者）所共有的特性。

## 36.3.4 估计方法

可以基于$\boldsymbol{y_i}$的边缘分布写出似然函数， 进行最大似然估计或限制最大似然估计(REML)。

最大似然估计实际是根据$\{\boldsymbol{y_i}\}$的联合分布进行的最大似然估计。 但是最大似然估计给出的方差、协方差不是无偏估计， 在小样本情形影响更大。

REML(Restricted Maximum Likelihood)是先用误差项独立同分布假设下的最小二乘方法估计出固定效应参数， 求出残差并利用残差估计方差结构， 这实际上是对因变量观测值向量进行了$\boldsymbol{n-p}$个线性独立的变换， 这些线性变换向量与设计阵的$\boldsymbol{p}$列正交， 这样可以得到方差、协方差的无偏估计； 再假定方差结构已知写出似然函数重新估计固定效应参数。 因为是分两步进行的估计， 所以两个用REML估计的嵌套模型， 仅固定效应相同时可比， 可以用REML得到的似然比检验比较其随机效应的差异。

另一方面， 如果两个嵌套模型的随机效应相同， 固定效应不同， 则可以用最大似然估计得到的似然比检验比较两个模型的固定效应的差异。

似然比检验计算精简模型和完全模型的对数似然函数值最大值(ML或REML)的差的$\boldsymbol{-2}$倍作为检验统计量， 在无显著差异的零假设下近似服从自由度等于参数个数之差的卡方分布， 统计量值超过右侧临界值时拒绝零假设。

## 36.3.5 ARMD数据建模

考虑对ARMD数据的分析。 主要的研究目标是考察处理组与对照组的差异， 以及视力随时间的变化。

#### 36.3.5.1 主效应和随机截距项

先考虑如下的主效应模型：

$$\mathbf{VISUAL}_{it} = \beta_0 + \beta_1 \times \mathbf{VISUAL0}_i + \beta_2 \times \mathbf{TIME}_{it} + \beta_3 \times \mathbf{TREAT}_i + b_{0i} + \varepsilon_{it},$$

这里$i$是病人编码， $t$是视力测量的时间序号(1,2,3,4)， $\mathbf{VISUAL}_{it}$是第$i$号病人在第$t$次随访的视力测量值， $\beta_0$是固定效应中的截距项， $\mathbf{VISUAL0}_i$是第$i$号病人在随机化分组前的视力的基线测量值， $\beta_1$属于固定效应， $\mathbf{TIME}_{it}$是$\mathbf{VISUAL}_{it}$的测量时间， 取值于$4, 12, 24, 52$周， $\beta_2$属于固定效应； $\mathbf{TREAT}_i$是处理组的示性函数， 第$i$病人属于处理组时为1， 处于对照组时为0， $\beta_3$属于固定效应。 $b_{0i}$是第$i$病人特有的截距项修正， 期望为0， 是随机效应， 此项存在使得同一病人的各观测之间有一个相等的正相关系数。 $\varepsilon_{it}$是随机误差项。

固定效应$\beta_3$体现了处理组与对照组的均值差异， 此项等于零的检验如果显著， 说明两组之间有显著差异。

因为处理组和对照组关于时间$\mathbf{TIME}_{it}$的斜率相等， 都是$\beta_2$， 所以这个模型的固定效应部分是一个平行线模型。

设$\varepsilon_{ij}$独立同分布N(0, $\sigma_e^2$)， $b_i$独立同分布N(0, $\sigma_b^2$)， $\{\varepsilon_{ij}\}$与$\{b_i\}$相互独立， 则

$$\begin{aligned}
\mathrm{Var}(\mathbf{VISUAL}_{it}) &= \sigma_b^2 + \sigma_e^2, \\
\mathrm{Cov}(\mathbf{VISUAL}_{it}, \mathbf{VISUAL}_{it'}) &= \sigma_b^2, (t \neq t') \\
\mathrm{Corr}(\mathbf{VISUAL}_{it}, \mathbf{VISUAL}_{it'}) &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2},
\end{aligned}$$

即同一受试者不同时间的测量值之间的相关系数为常数， 称这样的方差结构为复合对称(compound symmetry)。

用R程序计算如下：

```
library(nlme)
data(armd, package="nlmeU")
lme1 <- lme(
  visual ~ visual0 + time + treat.f,
  random = ~ 1 | subject,
  data = armd)
summary(lme1)
```

```
## Linear mixed-effects model fit by REML
##   Data: armd
##        AIC      BIC    logLik
##   6587.202 6615.764 -3287.601
##
## Random effects:
##  Formula: ~1 | subject
##         (Intercept) Residual
## StdDev:    8.967864 8.637692
##
## Fixed effects:  visual ~ visual0 + time + treat.f
##                   Value Std.Error  DF    t-value p-value
## (Intercept)    9.788402 2.6586549 632   3.681712  0.0003
## visual0        0.826563 0.0446365 231  18.517678  0.0000
## time          -0.235355 0.0167632 632 -14.039982  0.0000
## treat.fActive -3.476393 1.3184773 231  -2.636673  0.0089
##  Correlation:
##               (Intr) visul0 time
## visual0       -0.928
## time          -0.136 -0.002
## treat.fActive -0.269  0.026  0.015
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.10527500 -0.39197459  0.03661312  0.54464543  2.92150255
##
## Number of Observations: 867
## Number of Groups: 234
```

程序中 `random =` 是随机效应的设置， `| subject` 表明按照病人编码分组， 各组之间独立， 每组内部有随机效应， `~ 1` 表示随机效应中仅有一个随机的截距项（$b_{0i}$）。

结果中对 `treat.fActive` 的检验就是关于$H_0 : \beta_3 = 0$的检验， p值为0.0087，在0.05水平下显著， 两组之间有显著差异。 但处理组的视力更差。

输出结果中随机效应部分， 给出了随机截距项$b_i$的方差估计$\hat{\sigma}_b^2 = 8.97^2$, 随机误差$\epsilon_{ij}$的方差估计$\hat{\sigma}_e^2 = 8.64^2$。

结果给出了固定效应参数估计$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$之间的相关系数的估计， 如$\hat{\beta}_1$和$\hat{\beta}_2$的相关系数估计为$-0.002$。

#### 36.3.5.2 固定效应中不同斜率

处理组和对照组的视力随时间的变化的斜率不一定相等， 所以考虑如下的带有交叉项的模型：

$$\begin{aligned}\mathbf{VISUAL}_{it} =& \beta_0 + \beta_1 \times \mathbf{VISUAL0}_i + \beta_2 \times \mathbf{TIME}_{it} + \beta_3 \times \mathbf{TREAT}_i \\ &+ \beta_4 \times \mathbf{TIME}_{it} \times \mathbf{TREAT}_i + b_{0i} + \varepsilon_{it},\end{aligned}$$

这样，处理组关于$\mathbf{TIME}_{it}$的斜率为$\beta_2 + \beta_4$， 对照组则为$\beta_2$。

用R程序计算如下：

```
lme2 <- update(lme1, . ~ . + treat.f:time)
summary(lme2)
```

或直接写成：

```
lme2 <- lme(
  visual ~ visual0 + time + treat.f + treat.f:time,
  random = ~ 1 | subject,
  data = armd)
summary(lme2)
```

```
## Linear mixed-effects model fit by REML
##   Data: armd
##        AIC      BIC    logLik
##   6591.971 6625.286 -3288.986
##
## Random effects:
##  Formula: ~1 | subject
##         (Intercept) Residual
## StdDev:    8.978212 8.627514
##
## Fixed effects:  visual ~ visual0 + time + treat.f + treat.f:time
##                         Value Std.Error  DF   t-value p-value
## (Intercept)          9.288078 2.6818888 631  3.463260  0.0006
## visual0              0.826440 0.0446670 231 18.502244  0.0000
## time                -0.212216 0.0229295 631 -9.255150  0.0000
## treat.fActive       -2.422000 1.4999667 231 -1.614703  0.1077
## time:treat.fActive -0.049591 0.0335617 631 -1.477594  0.1400
##  Correlation:
##                     (Intr) visul0 time   trt.fA
## visual0             -0.920
## time                -0.185 -0.003
## treat.fActive       -0.295  0.022  0.335
## time:treat.fActive  0.126  0.002 -0.683 -0.476
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.18750513 -0.39692515  0.03204783  0.55138252  2.95132118
##
## Number of Observations: 867
## Number of Groups: 234
```

默认的参数估计方法是REML。

这里的检验是第三类检验， 即有某项的模型与仅排除该项的模型的比较。 用 `anova()` 则选择进行第一类检验， 即依次增加项的检验：

```
anova(lme2)
```

```
##              numDF denDF  F-value p-value
## (Intercept)      1   631 5399.285  <.0001
## visual0          1   231  343.825  <.0001
## time             1   631  196.499  <.0001
## treat.f          1   231    6.942   0.009
## time:treat.f     1   631    2.183   0.140
```

因为交叉项 `time:treat.fActive` 在最后面， 所以其第三类检验和第一类检验相同。

可以用 `anova()` 比较两个嵌套模型， 为了比较随机效应相同、固定效应不同的模型， 需要两个模型都使用最大似然估计。 如：

```
lme1.ml <- update(lme1, method="ML")
lme2.ml <- update(lme2, method="ML")
anova(lme1.ml, lme2.ml)
```

```
##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme1.ml     1  6 6579.854 6608.444 -3283.927
## lme2.ml     2  7 6579.672 6613.027 -3282.836 1 vs 2 2.181979  0.1396
```

取不同斜率有一定作用。

### 36.3.5.3 随机斜率

考虑每个人的视力随时间变化分别有不同的斜率。 模型公式为

$$\begin{aligned}\mathrm{VISUAL}_{it} =& \beta_0 + \beta_1 \times \mathrm{VISUAL0}_i + \beta_2 \times \mathrm{TIME}_{it} + \beta_3 \times \mathrm{TREAT}_i \\ &+ \beta_4 \times \mathrm{TIME}_{it} \times \mathrm{TREAT}_i + b_{0i} + b_{1i}\mathrm{TIME}_{it} + \varepsilon_{it}.\end{aligned}$$

用R程序计算如下：

```
lme3 <- update(lme2, random = ~ 1 + time | subject)
summary(lme3)
```

```
## Linear mixed-effects model fit by REML
##   Data: armd
##        AIC      BIC    logLik
##   6453.824 6496.657 -3217.912
##
## Random effects:
##  Formula: ~1 + time | subject
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 7.0666328 (Intr)
## time        0.2729987 0.143
## Residual    6.7442472
##
## Fixed effects:  visual ~ visual0 + time + treat.f + treat.f:time
##                        Value Std.Error  DF   t-value p-value
## (Intercept)         4.768536 2.3195550 631  2.055798  0.0402
## visual0             0.908896 0.0391891 231 23.192601  0.0000
## time               -0.215285 0.0316627 631 -6.799343  0.0000
## treat.fActive      -2.290098 1.1835442 231 -1.934949  0.0542
## time:treat.fActive -0.056374 0.0462155 631 -1.219803  0.2230
##  Correlation:
##                    (Intr) visul0 time   trt.fA
## visual0            -0.934
## time               -0.073  0.003
## treat.fActive      -0.274  0.026  0.138
## time:treat.fActive  0.050 -0.002 -0.685 -0.207
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.93038550 -0.33999935  0.04188037  0.46186462  2.98622721
##
## Number of Observations: 867
## Number of Groups: 234
```

这里程序中 `random` 参数的 `~ 1 + time` 中的 `time` 表示每个病人关于 `time` 自变量有单独的斜率项， 此斜率项为均值等于零的随机变量。

`lme2` 和 `lme3` 的固定效应相同， 仅随机效应有差别， 当固定效应相同时用 `anova()` 函数比较两个模型的随机效应部分， 应使用REML估计。程序如：

```
anova(lme2, lme3)
```

```
##      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme2     1  7 6591.971 6625.286 -3288.986
## lme3     2  9 6453.824 6496.657 -3217.912 1 vs 2 142.1469  <.0001
```

结果表明加入 `time` 的随机效应斜率项是有意义的。 从AIC结果来看也是 `lme3` 较好。

现在固定效应中的交叉项显著性不高，去掉此项：

```
lme4 <- update(lme3, . ~ . - treat.f:time)
summary(lme4)
```

```
## Linear mixed-effects model fit by REML
##   Data: armd
##        AIC      BIC  logLik
##   6448.999 6487.083 -3216.5
##
## Random effects:
##  Formula: ~1 + time | subject
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 7.0799183 (Intr)
## time        0.2734706 0.138
## Residual    6.7423971
##
## Fixed effects:  visual ~ visual0 + time + treat.f
##                   Value Std.Error  DF    t-value p-value
## (Intercept)    4.927439 2.3181699 632   2.125573  0.0339
## visual0        0.908509 0.0392129 231  23.168651  0.0000
## time          -0.241733 0.0230894 632 -10.469443  0.0000
## treat.fActive -2.593006 1.1585178 231  -2.238210  0.0262
##  Correlation:
##               (Intr) visul0 time
## visual0       -0.935
## time          -0.054  0.003
## treat.fActive -0.270  0.026 -0.006
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.94548685 -0.34966982  0.04283532  0.45701780  2.95567436
##
## Number of Observations: 867
## Number of Groups: 234
```

从AIC来看此模型较优。

lme3和lme4随机效应相同而固定效应不同， 为了用 `anova()` 检验固定效应中交叉项， 应使用最大似然估计：

```
lme3.ml <- update(lme3, method="ML")
lme4.ml <- update(lme4, method="ML")
anova(lme3.ml, lme4.ml)
```

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lme3.ml      1  9 6442.015 6484.900 -3212.008
## lme4.ml      2  8 6441.515 6479.635 -3212.757 1 vs 2 1.499852  0.2207
```

可见固定效应中处理组和对照组在time上的不同斜率是不必要的， 关于time可以使用平行线模型。

## 36.3.6 Orthodont数据建模

nlme包中有一个例子数据Orthodont， 是27名儿童（16名男生，11名女生）在8、10、12、14岁测量的某个口腔距离（脑下垂体到翼上颌裂的距离）。 这是纵向数据， 同一名儿童的不同年龄的观测是相关的； 这27名儿童看成是从一个很大的总体中抽样得到的， 并不关系每一个儿童具体的情况。

```
library(nlme)
```

```
##
## 载入程辑包：'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
data(Orthodont, package = "nlme")
```

### 36.3.6.1 主效应和误差项等方差

考虑年龄增长时距离增长， 并考虑性别的差异。 用平行线模型。

```r
orth.lme1 <- lme(
  distance ~ age + Sex,
  random = ~ 1 | Subject,
  data=Orthodont)
summary(orth.lme1)
```

```
## Linear mixed-effects model fit by REML
##   Data: Orthodont
##        AIC      BIC    logLik
##   447.5125 460.7823 -218.7563
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    1.807425 1.431592
##
## Fixed effects:  distance ~ age + Sex
##                 Value Std.Error DF   t-value p-value
## (Intercept) 17.706713 0.8339225 80 21.233044  0.0000
## age          0.660185 0.0616059 80 10.716263  0.0000
## SexFemale   -2.321023 0.7614168 25 -3.048294  0.0054
##  Correlation:
##           (Intr) age
## age       -0.813
## SexFemale -0.372  0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.74889609 -0.55034466 -0.02516628  0.45341781  3.65746539
##
## Number of Observations: 108
## Number of Groups: 27
```

年龄和性别的固定效应都显著。

### 36.3.6.2 加入交互作用

增加年龄和性别的交互作用效应， 即不同性别使用不同的年龄斜率项：

```
orth.lme2 <- update(orth.lme1,
  . ~ . + Sex:age)
summary(orth.lme2)
```

```
## Linear mixed-effects model fit by REML
##   Data: Orthodont
##        AIC      BIC    logLik
##   445.7572 461.6236 -216.8786
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    1.816214 1.386382
##
## Fixed effects:  distance ~ age + Sex + age:Sex
##                   Value Std.Error DF   t-value p-value
## (Intercept)   16.340625 0.9813122 79 16.651810  0.0000
## age            0.784375 0.0775011 79 10.120823  0.0000
## SexFemale      1.032102 1.5374208 25  0.671321  0.5082
## age:SexFemale -0.304830 0.1214209 79 -2.510520  0.0141
##  Correlation:
##               (Intr) age    SexFml
## age           -0.869
## SexFemale     -0.638  0.555
## age:SexFemale  0.555 -0.638 -0.869
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.59804400 -0.45461690  0.01578365  0.50244658  3.68620792
##
## Number of Observations: 108
## Number of Groups: 27
```

orth.lme1和orth.lme2方差结构相同， 固定效应不同， 需要用最大似然估计对应的似然比检验进行比较：

```
orth.lme1b <- update(orth.lme1, method = "ML")
orth.lme2b <- update(orth.lme2, method = "ML")
anova(orth.lme1b, orth.lme2b)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## orth.lme1b     1  5 444.8565 458.2671 -217.4282
## orth.lme2b     2  6 440.6391 456.7318 -214.3195 1 vs 2 6.217427  0.0126
```

不同性别使用不同的age斜率项是有意义的。 因为age和SexFemale的交互项系数为负值且显著， 说明女性的age斜率项显著地低于男性， 即女生的因变量随年龄的增长率低于男生。

### 36.3.6.3 不同性别的误差项使用不同方差

如果将男、女分两组分别建模， 可以发现其随机误差部分的方差估计相差较大。 如下的程序可以设定男、女两组的误差项使用不同的方差， 而随机截距项仍假设相同的方差：

```
orth.lme3 <- update(orth.lme2,
  weights = varIdent(form = ~ 1 | Sex))
summary(orth.lme3)
```

也可以直接写成

```
orth.lme3 <- lme(
  distance ~ age + Sex + Sex:age,
  random = ~ 1 | Subject,
  weights = varIdent(form = ~ 1 | Sex),
  data = Orthodont)
summary(orth.lme3)
```

```
## Linear mixed-effects model fit by REML
##   Data: Orthodont
##        AIC      BIC    logLik
##   429.2205 447.7312 -207.6102
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:     1.84757 1.669823
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Sex
##  Parameter estimates:
##      Male    Female
## 1.0000000 0.4678944
## Fixed effects:  distance ~ age + Sex + Sex:age
##                   Value Std.Error DF   t-value p-value
## (Intercept)   16.340625 1.1450945 79 14.270111  0.0000
## age            0.784375 0.0933459 79  8.402883  0.0000
## SexFemale      1.032102 1.4039842 25  0.735124  0.4691
## age:SexFemale -0.304830 0.1071828 79 -2.844016  0.0057
##  Correlation:
##               (Intr) age    SexFml
## age           -0.897
## SexFemale     -0.816  0.731
## age:SexFemale  0.781 -0.871 -0.840
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.00556474 -0.63419474  0.01890475  0.55016878  3.06446971
##
## Number of Observations: 108
## Number of Groups: 27
```

程序中 `weights` 参数用来指定非独立同分布情形的误差项方差结构， 其中 `varIdent()` 用来规定在某个分类变量的每一类中独立且方差相同的方差结构， 这里 `~ 1 | Sex` 则指定了按性别分成两组， 每一组内的误差项假定是相互独立且方差相等的， 两组的误差项方差不相等。

在输出结果中， 显示了不同层(stratum)， 这里是男性和女性两组的误差项方差比例关系， 比例为$1:0.47$。

lme2和lme3的固定效应相同， 方差结构不同且构成嵌套关系， 可以用基于REML的似然比检验进行比较：

```
anova(orth.lme2, orth.lme3)
```

```
##           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## orth.lme2     1  6 445.7572 461.6236 -216.8786
## orth.lme3     2  7 429.2205 447.7312 -207.6102 1 vs 2 18.53677  <.0001
```

两个模型有显著差异， 说明男性组与女性组使用不同的误差项方差是必要的。

#### 36.3.6.4 删去不同性别的不同截距

上面结果中Sex的主效应不显著，可以删去此主效应， 使得固定效应中男女的截距项相同而斜率项不同。 程序如下：

```
orth.lme4 <- update(orth.lme3,
  . ~ . - Sex)
summary(orth.lme4)
```

```
## Linear mixed-effects model fit by REML
##   Data: Orthodont
##        AIC      BIC    logLik
##   430.2763 446.2001 -209.1381
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    1.842395 1.665563
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Sex
##  Parameter estimates:
##      Male    Female
## 1.0000000 0.4685187
## Fixed effects:  distance ~ age + age:Sex
##                   Value Std.Error DF   t-value p-value
## (Intercept)   17.026899 0.6611240 79 25.754472   0e+00
## age            0.734206 0.0635120 79 11.560109   0e+00
## age:SexFemale -0.238642 0.0580447 79 -4.111355   1e-04
##  Correlation:
##               (Intr) age
## age           -0.761
## age:SexFemale  0.305 -0.693
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.1173385 -0.6522050  0.0218480  0.5116131  3.0887448
##
## Number of Observations: 108
## Number of Groups: 27
```

比较两个模型：

```r
orth.lme3b <- update(orth.lme3, method="ML")
orth.lme4b <- update(orth.lme4, method="ML")
anova(orth.lme3b, orth.lme4b)
```

```
##            Model df      AIC      BIC    logLik   Test   L.Ratio p-value
## orth.lme3b     1  7 423.3524 442.1273 -204.6762
## orth.lme4b     2  6 421.9128 438.0056 -204.9564 1 vs 2 0.5604856  0.4541
```

可见两个模型没有显著差异。

## 36.4 非线性混合模型介绍

待完成。

$$a$$

## References

Andrzej Galecki, Tomasz Burzykowski. 2013. *Linear Mixed-Effects Models Using r*. Springer.