BI476: Biostatistics - Case Studies

Lec06C: Clustered and Longitudinal Data Analysis

Maoying,Wu
ricket.woo@gmail.com

Dept. of Bioinformatics & Biostatistics
Shanghai Jiao Tong University

Spring, 2018

Outline

1. Clustered and Longitudinal Data Analysis
   - Exploratory Analysis
   - Inference Analysis

Next Section ...

1. Clustered and Longitudinal Data Analysis
   - Exploratory Analysis
   - Inference Analysis

The assumption of independence is often deviated

- In all the models considered so far the outcomes Y_i's are assumed to be indepdent.
- However, this does not hold in most of the situations:
  - **Longitudinal data**: repeated measures over time on the same subjects;
  - **Clustered data**: measurements on related subjects;
- Modeling approaches:
  - **Repeated measures** and **generalized estimating equations (GEEs)** to explicitly model the correlation structure.
  - **Multi-level modeling** to consider the hierarchical structure of the study design.

Exploratory Data Analysis

CD4+ Data

A total of 2376 observations of CD4+ cell counts with respective time since seroconversion (detectable HIV antibodies) for 369 infected men enrolled in the Multicenter AIDS Cohort Study (MACS).

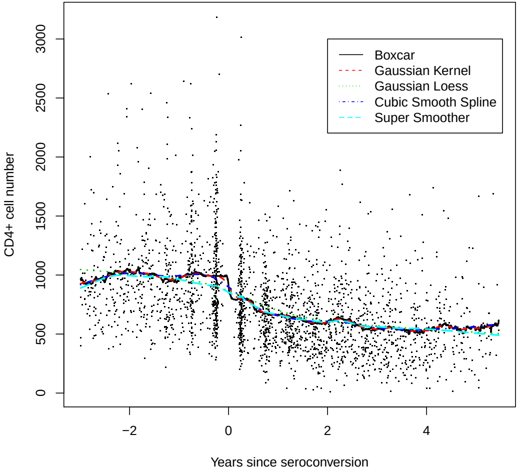
```
##      Time CD4  Age Packs Drugs Sex Cesd    ID
## 1 -0.742 548 6.57     0     0   5    8 10002
## 2 -0.246 893 6.57     0     1   5    2 10002
## 3  0.244 657 6.57     0     1   5   -1 10002
## 4 -2.730 464 6.95     0     1   5    4 10005
## 5 -2.251 845 6.95     0     1   5   -4 10005
## 6 -0.222 752 6.95     0     1   5   -5 10005
```

- Scatterplot with smoothers.
- "Spaghetti" plot.
- Exploring the correlation structure within each cluster.

1. Scatterplots with smoothers

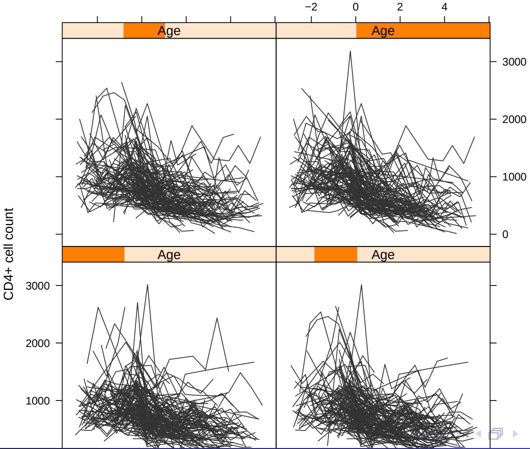
```r
# Draw the scatterplot
plot(CD4 ~ Time, data=cd4, pch=".", xlab="Years since seroconversion",
        ylab="CD4+ cell number")
# Draw the smoother curve
with(cd4, {
        lines(ksmooth(Time, CD4, kernel="box"), lty=1, col=1, lwd=2)
        lines(ksmooth(Time, CD4, kernel="normal"), lty=2, col=2, lwd=2)
        lines(loess.smooth(Time, CD4, family="gaussian"), lty=3, col=3,
                lwd=2)
        lines(smooth.spline(Time, CD4), lty=4, col=4, lwd=2)
        lines(supsmu(Time, CD4), lty=5, col=5, lwd=2)
})
legend(2,3000,legend=c("Boxcar", "Gaussian Kernel", "Gaussian Loess",
                "Cubic Smooth Spline", "Super Smoother"), lty=1:5,
        col=1:5)
```

1. Scatterplots with smoothers



2. Variation accross each individual

```
library(lattice)
xyplot(CD4 ~ Time | equal.count(Age, 4), data=cd4, type="l", group=ID,
       xlab="Years since seroconversion", col.line="gray20",
       ylab="CD4+ cell count", strip=strip.custom(var.name="Age"))
```

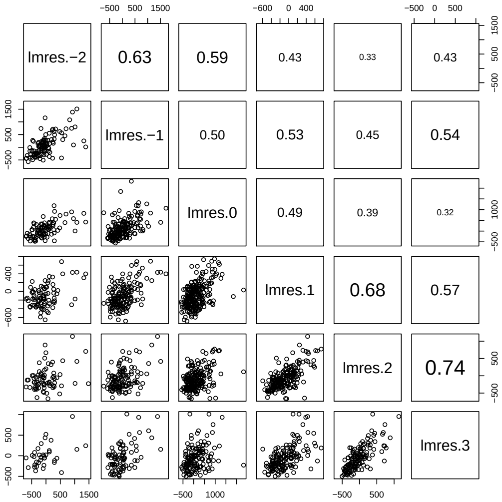


3.Correlation structure

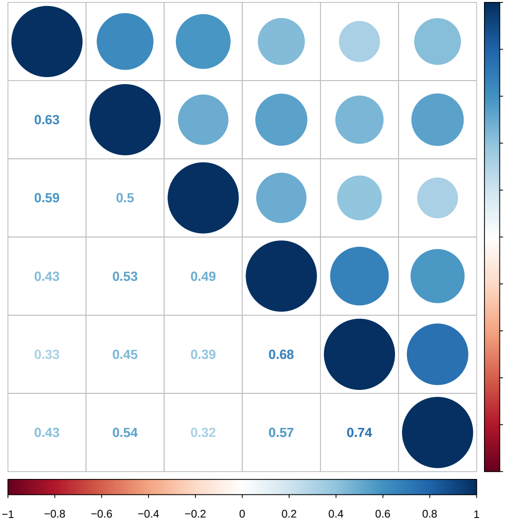
```r
# Fit a linear model using the pooled data
CD4.lm <- lm(CD4 ~ Time, data=cd4)
# Obtain the Pearson's residuals
cd4$lmres <- resid(CD4.lm)
# Round time to integer
cd4$roundyr <- round(cd4$Time)
# Reshape the data
cd4w <- reshape(cd4[,c("ID", "lmres", "roundyr")],
        direction="wide", v.names="lmres", timevar="roundyr",
        idvar="ID")

panel.cor <- function(x, y, digits=2, prefix="", cex.cor){
        usr <- par("usr"); on.exit(par(usr))
        par(usr=c(0,1,0,1))
        r <- abs(cor(x, y, use="pairwise.complete.obs"))
        txt <- format(c(r, 0.123456789),digits=digits)[1]
        txt <- paste(prefix, txt, sep="")
        if (missing(cex.cor)) cex <- 0.8/strwidth(txt)
        text(0.5, 0.5, txt, cex=cex*r)
}
pairs(cd4w[,c(5,2,3,6:8)], upper.panel = panel.cor)
```

3.Correlation structure



3.Correlation structure by bubble plot



1. Pooled Analysis

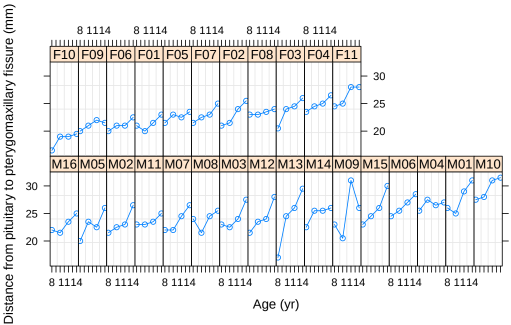
$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$$

where

- $\epsilon_{ij} \sim N(0, \sigma^2)$

Orthodont Data

```
library(nlme, quietly=TRUE)
data(Orthodont)
plot(Orthodont, layout=c(16, 2))
```



Pooled Analysis

```
lm1 <- lm(distance ~ I(age-1)*Sex, data=Orthodont)
summary(lm1)
```

```
##
## Call:
## lm(formula = distance ~ I(age - 1) * Sex, data = Orthodont)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.616 -1.322 -0.168  1.330  5.247
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       17.4886     1.0127   17.27  < 2e-16 ***
## I(age - 1)         0.6320     0.0988    6.39  4.7e-09 ***
## Sex1              -0.3636     1.0127   -0.36     0.72
## I(age - 1):Sex1    0.1524     0.0988    1.54     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.26 on 104 degrees of freedom
## Multiple R-squared:  0.423,Adjusted R-squared:  0.406
## F-statistic: 25.4 on 3 and 104 DF,  p-value: 2.11e-12
```

2. Data Reduction Analysis

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \epsilon_{ij}$$

where

- $\epsilon_{ij} \sim N(0, \sigma^2)$

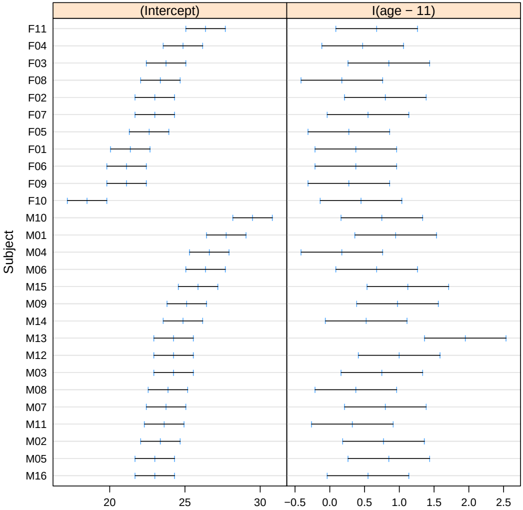
Data Reduction Analysis

```
fm <- nlme::lmList(distance ~ I(age-11) | Subject, data=Orthodont)
summary(fm)
```

```
## Call:
##   Model: distance ~ I(age - 11) | Subject
##    Data: Orthodont
##
## Coefficients:
##    (Intercept)
##     Estimate Std. Error t value Pr(>|t|)
## M16     23.0      0.655    35.1 7.23e-39
## M05     23.0      0.655    35.1 7.23e-39
## M02     23.4      0.655    35.7 3.13e-39
## M11     23.6      0.655    36.1 1.80e-39
## M07     23.8      0.655    36.3 1.37e-39
## M08     23.9      0.655    36.4 1.04e-39
## M03     24.2      0.655    37.0 4.64e-40
## M12     24.2      0.655    37.0 4.64e-40
## M13     24.2      0.655    37.0 4.64e-40
## M14     24.9      0.655    38.0 1.23e-40
## M09     25.1      0.655    38.4 7.33e-41
## M15     25.9      0.655    39.5 1.58e-41
## M06     26.4      0.655    40.3 5.81e-42
```

Data Reduction Analysis

```r
plot(intervals(fm))
```



3. Generalized Estimating Equation (GEE)

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}, \mathbf{y}_i \in \mathbb{R}^{n_i}$$

A normal linear model for $\mathbf{y}$ is

$$E(\mathbf{y}) = \mathbf{X}\beta = \mu; \mathbf{y} \sim \text{MVN}(\mu, \mathbf{V})$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

where $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ design matrix for unit i, and β is a parameter vector of length p.

GEE: Variance-Covariance Matrix

The variance-covariance matrix for measurements for unit i is

$$\mathbf{V}_i = \begin{bmatrix} \sigma_{i11} & \sigma_{i12} & \dots & \sigma_{i1n_i} \\ \sigma_{i21} & \sigma_{i22} & \dots & \sigma_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{in_i1} & \sigma_{in_i2} & \dots & \sigma_{in_in_i} \end{bmatrix}$$

and the overall variance-covariance matrix has the block diagonal form:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \mathbf{V}_2 & \dots & \mathbf{O} \\ \vdots & \vdots & \ddots & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{V}_N \end{bmatrix}$$

GEE: Estimating the coefficient

If $\mathbf{V}_i$ is known, the maximum likelihood estimator is obtained by solving the score equations:

$$\mathbf{U}(\beta) = \frac{\partial l}{\partial \beta} = \mathbf{X^T V^{-1}}(\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^{N} \mathbf{X_i^T V_i^{-1}}(\mathbf{y_i} - \mathbf{X_i}\beta) = \mathbf{0}$$

where l is the log-likelihood function. The solution is

$$\hat{\beta} = (\mathbf{X^T V^{-1} X})^{-1} \mathbf{X^T V^{-1} y} = (\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i^T V_i^{-1} X_i})^{-1} (\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i^T V_i^{-1} y_i})$$

with

$$\text{Var}(\hat{\beta}) = (\mathbf{X^T V^{-1} X})^{-1} = (\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i^T V_i^{-1} X})^{-1}$$

and $\hat{\beta}$ is asymptotically normal.

Alternate Estimator of $\hat{\beta}$ and $\hat{\mathbf{V}}$

In practice, $\mathbf{V}$ is usually unknown and has to be estimated from the data by an iterative approach:

1. Starting with an intial $\mathbf{V}$ (e.g., the identity matrix)
2. Calculating an estiamte $\hat{\beta}$ and hence the linear predictors $\hat{\mu} = \mathbf{X}\hat{\beta}$ and the residuals $\mathbf{r} = \mathbf{y} - \hat{\mu}$
3. Computing $\mathbf{V}$
4. Repeating the process until convergence is achieved.

An alternative estimator of $\hat{\mathbf{V}}$

The above approach will underestimate the variance of $\hat{\beta}$. Therefore, a preferable alternative for $\hat{\mathbf{V}}$ is

$$\mathbf{V}_s(\hat{\beta}) = \mathcal{J}^{-1}\mathbf{C}\mathcal{J}^{-1}$$

where

$$\mathcal{J} = \mathbf{X^T}\hat{\mathbf{V}}^{-1}\mathbf{X} = \sum_{i=1}^{N} \mathbf{X}_i^T \hat{V}_i^{-1}\mathbf{X}_i$$

and

$$\mathbf{C} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i^T}\hat{\mathbf{V}}_i^{-1}(\mathbf{y_i} - \mathbf{X_i}\hat{\beta})(\mathbf{y}_i - \mathbf{X}_i\hat{\beta})^T\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i.$$

- $\mathbf{V_s}(\beta)$ is called the **information sandwich estimator**
- It is also called the **Huber estimator**.
- Consistent estimator of Var$(\hat{\beta})$ when $\mathbf{V}$ is unknown
- Robust to misspecification of $\mathbf{V}$

Chocies of $\mathbf{V}_i$

The within-unit (cluster) variance-covariance matrix $\mathbf{V}_i$ has some choices:

- Independent model
- Exchangeable/spherical model
- Autoregressive model
- Unstructured correlation model
- User-defined model

Independent model

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Exchangeable model

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

Autoregressive model

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{bmatrix}$$

Unstructured model

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{bmatrix}$$

Repeated Measures Models for Non-Normal Data

For the generalized linear model:

$$E(Y_i) = \mu_i, g(\mu_i) = \mathbf{x}_i^T \beta = \eta_i$$

For repeated measures, let

- $\mathbf{y}_i$: the vector of responses for unit i with $E(\mathbf{y}_i) = \mu_i, g(\mu_i) = \mathbf{X}_i^T \beta$
- $\mathbf{D}_i$: the matrix of derivatives $\partial\mu_i/\partial\beta_j$
- The **GEE analogue** of score equations are:

$$\mathbf{U} = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{N}} \mathbf{D}_{\mathbf{i}}^{\mathbf{T}} \mathbf{V}_{\mathbf{i}}^{-\mathbf{1}} (\mathbf{y}_{\mathbf{i}} - \mu_{\mathbf{i}}) = \mathbf{0}$$

  which can be called the **quasi-score equations**.
- The matrix $\mathbf{V}_i$ can be written as

$$\mathbf{V}_{\mathbf{i}} = \mathbf{A}_{\mathbf{i}}^{\mathbf{1/2}} \mathbf{R}_{\mathbf{i}} \mathbf{A}_{\mathbf{i}}^{\mathbf{1/2}} \phi$$

  where
  - $\mathbf{A}_i$: the diagonal matrix with elements var(y_{ik})
  - $\mathbf{R}_i$: correlation matrix for $\mathbf{y}_i$
  - ϕ: constant for overdispersion parameter

GEEs: Summary

- If $\mathbf{R}_i$ are correctly specified, the estimator $\hat{\beta}$ is consistent and asymptotically normal.
- $\hat{\beta}$ is fairly robust against mis-specification of $\mathbf{R}_i$
- Knowledge of the study design and results from exploratory analysis should be used to select a plausible form of $\mathbf{R}_i$.
- Use a small number of parameters (exchangeable or autoregressive correlation)
- Use **sandwich estimator** for var$(\hat{\beta})$

Inference

1. Start with $\mathbf{R}_i = I_n$ and $\phi = 1$
2. Obtain the parameter $\hat{\beta}$
3. Calculate the fitted values $\hat{\mu}_i = g^{-1}(\mathbf{X}_\mathbf{i}^\mathbf{T}\beta)$ and the residuals $\mathbf{y}_i - \hat{\mu}_i$
4. Estimate $\mathbf{A}_i, \mathbf{R}_i$ and ϕ
5. Repeat the above process.

GEEs in R

```
geepack::geeglm(formula, family=gaussian, data, id, zcor=
    NULL,
        constr, std.err="san.se")
```

`formula`	Symbolic description of the model to be fitted.
`family`	Description of the error distribution and link function.
`data`	Optional dataframe
`id`	Vector that identifies the clusters
`zcor`	User-defined correlation structure.
`constr`	Working correlation structure: "ind", "ex", "ar1", "unstructured", "use
`std.err`	Type of standard error to be calculated, "san.se" for robust sandwic

GEE Modeling (1)

```
library(geepack, quietly=TRUE)
gee1 <- geeglm(distance ~ I(age-11)*Sex, data=Orthodont, id=Subject, corst
summary(gee1)
```

```
##
## Call:
## geeglm(formula = distance ~ I(age - 11) * Sex, data = Orthodont,
##     id = Subject, corstr = "exchangeable")
##
##  Coefficients:
##                     Estimate Std.err    Wald Pr(>|W|)
## (Intercept)          23.8082  0.3749 4033.27   <2e-16 ***
## I(age - 11)           0.6320  0.0584  116.96   <2e-16 ***
## Sex1                  1.1605  0.3749    9.58   0.0020 **
## I(age - 11):Sex1      0.1524  0.0584    6.80   0.0091 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)     4.91    1.01
##
## Correlation: Structure = exchangeable  Link = identity
```

4. Multilevel models: random intercept model

Y_{jk} is the response of the k-th subject in the j-th cluster.

$$Y_{jk} = \mu + a_j + e_{jk}$$

- $a_j \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$;
- $e_{jk} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$
- $a_j \perp e_{jk}$

then we can get:

- $E(Y_{jk}) = \mu$, $\text{var}(Y_{jk}) = \sigma_a^2 + \sigma_e^2$.
- $\text{cov}(Y_{jk}, Y_{jm}) = \sigma_a^2$.
- $\text{cov}(Y_{jk}, Y_{lm}) = 0$.

In this model,

- μ is a **fixed effect.**
- a_j is a **random effect.**
- **mixed model**
- The parameters of interest are μ, σ_a^2 and σ_e^2.

4. Multilevel models: random slope/intercept model

Y_{jk} is the measurement at t_k on subject j.

$$Y_{jk} = \beta_0 + a_j + (\beta_1 + b_j)t_k + e_{jk}$$

- β_0, β_1 are the population-level intercept and slope parameters.
- $a_j \overset{\text{iid}}{\sim} N(0, \sigma_a^2)$ is the difference from β_0 specific to subject j.
- $b_j \overset{\text{iid}}{\sim} N(0, \sigma_b^2)$ is the difference from β_1 specific to subject j.
- $e_{jk} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$ is the random error.
- a_j, b_j and e_{jk} are all assumed to be independent.

then we can get:

- $E(Y_{jk}) = \beta_0 + \beta_1 t_k$, $\text{var}(Y_{jk}) = \sigma_a^2 + t_k^2\sigma_b^2 + \sigma_e^2$.
- $\text{cov}(Y_{jk}, Y_{jm}) = \sigma_a^2 + t_k t_m \sigma_b^2$.
- $\text{cov}(Y_{jk}, Y_{lm}) = 0$.

In this model,

- β_0, β_1 is a **fixed effect**.
- a_j, b_j is a **random effect**.

General Mixed Models with Normal Response

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where

- $\beta \in \mathbb{R}^p$ are the fixed effects.
- $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{e} \in \mathbb{R}^n$ are the random effects.
- $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{Z} \in \mathbb{R}^{n \times q}$ are design matrices.
- $\mathbf{u}$ and $\mathbf{e}$ are assumed to be normally distributed.
- $E(\mathbf{y}) = \mathbf{X}\beta$ summarizes the **non-random** component of the model.
- $\mathbf{Z}\mathbf{u}$ describes the between-subjects random effects.
- $\mathbf{e}$ describes the within-subjects random effects.
- $\mathbf{u} \sim N(0, \mathbf{G}), \mathbf{e} \sim N(0, \mathbf{R})$
- The variance-covariance matrix for $\mathbf{y}$:

$$\mathbf{V}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathbf{T}} + \mathbf{R}$$

- The parameters can be estimated using the methods of either **maximum likelihood (ML)** or **restricted maximum likelihood (REML)**.

The Linear Mixed Model and Generalized

- Estimation of fixed effects and variance parameters using maximum likelihood (ML) or restricted maximum likelihood (REML)
- Predition of the random effects using best prediction.

Generalized Response

- $f(\mathbf{y}|\mathbf{u}) = \exp[\mathbf{y}^{\mathbf{T}}(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - 1^T b(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) + 1^T c(\mathbf{y})]$
- $\mathbf{u} \sim N(0, \mathbf{G})$
- Estimation and predction

$$\begin{bmatrix} \hat{\beta} \\ \hat{G} \end{bmatrix} = \arg\max_{\beta,G} f(\mathbf{y}; \beta, \mathbf{G})$$

Generalized Linear Mixed Models in R

```
nlme::lme(fixed, data, random)
```

`fixed`	two-sided linear formula for the fixed-effects part of the mode
`data`	data.frame containing the variables in the model.
`random`	one-sided formula for the random-effects part of the model.
`method`	"REML" for maximizing restricted log-likelihood; "ML" for max mizing the log-likelihood.

GLMM modeling (1)

```
library(nlme)
# Random intercept
lme1 <- lme(distance ~ I(age-11)*Sex, random=~1|Subject,
                data=Orthodont)

# Random slope
lme2 <- lme(distance ~ I(age-11)*Sex, random=~I(age-11)-1|Subject,
                data=Orthodont)

# Random intercept, random slope
lme3 <- lme(distance ~ I(age-11)*Sex, random=~I(age-11)|Subject,
                data=Orthodont)
```

GLMM modeling (2): Fixed Effects

```
fixef(lme1)

##      (Intercept)       I(age - 11)              Sex1 I(age - 11):Sex1
##           23.808             0.632             1.161            0.152

fixef(lme2)

##      (Intercept)       I(age - 11)              Sex1 I(age - 11):Sex1
##           23.808             0.632             1.161            0.152

fixef(lme3)

##      (Intercept)       I(age - 11)              Sex1 I(age - 11):Sex1
##           23.808             0.632             1.161            0.152
```

GLMM modeling (3): Random Effects

```
VarCorr(lme1)
```

```
## Subject = pdLogChol(1)
##             Variance StdDev
## (Intercept) 3.30     1.82
## Residual    1.92     1.39
```

```
VarCorr(lme2)
```

```
## Subject = pdLogChol(I(age - 11) - 1)
##             Variance StdDev
## I(age - 11) 1.33e-09 3.64e-05
## Residual    5.09e+00 2.26e+00
```

```
VarCorr(lme3)
```

```
## Subject = pdLogChol(I(age - 11))
##             Variance StdDev Corr
## (Intercept) 3.3501   1.83   (Intr)
## I(age - 11) 0.0325   0.18   0.206
## Residual    1.7162   1.31
```